BanNERD: A Benchmark Dataset and Context-Driven Approach for Bangla Named Entity Recognition

Md. Motahar Mahtab[1], Faisal Ahamed Khan[1], Md. Ekramul Islam[1], Md. Shahad Mahmud Chowdhury[1], Labib Imam Chowdhury[1], Sadia Afrin[1], Hazrat Ali[1], Mohammad Mamun Or Rashid[2], Nabeel Mohammed[3], Mohammad Ruhul Amin[4],

[1]Giga Tech Limited, Dhaka, Bangladesh, [2]Bangladesh Computer Council, Dhaka, Bangladesh, [3]North South University, Dhaka, Bangladesh, [4]Fordham University, New York, USA,

Correspondence: faisal.cse06@gigatechltd.com

Abstract

In this study, we introduce **BanNERD**, the most extensive human-annotated and validated **B**angla **N**amed **E**ntity **R**ecognition **D**ataset to date, comprising over 85,000 sentences. BanNERD is curated from a diverse array of sources, spanning over 29 domains, thereby offering a comprehensive range of generalized contexts. To ensure the dataset's quality, expert linguists developed a detailed annotation guideline tailored to the Bangla language. All annotations underwent rigorous validation by a team of validators, with final labels being determined via majority voting, thereby ensuring the highest annotation quality and a high IAA score of 0.88. In a cross-dataset evaluation, models trained on BanNERD consistently outperformed those trained on four existing Bangla NER datasets. Additionally, we propose a method named *BanNERCEM* (Bangla NER context-ensemble method) which outperforms existing approaches on Bangla NER datasets and performs competitively on English datasets using lightweight Bangla pretrained LLMs. Our approach passes each context separately to the model instead of previous concatenation-based approaches achieving the highest average macro F1 score of 81.85% across 10 NER classes, outperforming previous approaches and ensuring better context utilization. We are making the code and datasets publicly available at `https://github.com/eblict-gigatech/BanNERD` in order to contribute to the further advancement of Bangla NLP.

1 Introduction

NER focuses on discerning and classifying mainly but not limited to proper nouns (Chinchor and Robinson, 1998) in texts, facilitating the extraction of crucial content elements like individuals, organizations, and locations. Despite substantial advancements in NER for resource-rich languages including English, German, French, etc. a significant gap persists for the Bangla language due to lack of quality resources and linguistic complexity.

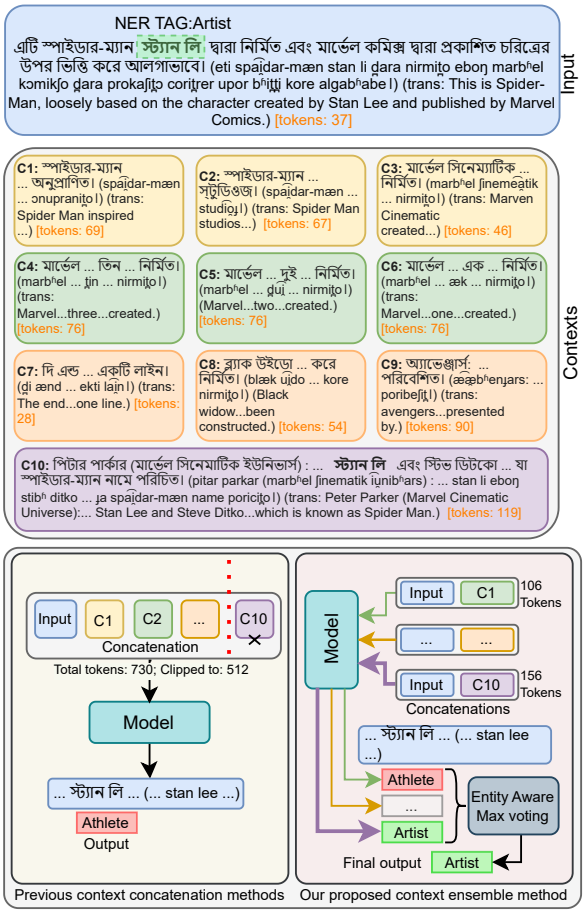

Figure 1: Our proposed *BanNERCEM* correctly identifies 'স্ট্যান লি' (ipa:stan li) (trans:Stan Lee) as 'Artist' instead of 'Athlete' predicted by Wang et al. (2022a). Out of ten contexts, 'Artist' got three votes and Athlete got 1 vote leading to a correct classification. Voting among contexts thus amplifies the chances of correct classification by choosing the NER tag most seen with the target entity (the target entity here is 'স্ট্যান লি' (ipa:stan li) (trans:Stan Lee).

Due to the high degree of inflection in Bangla, the same word can exhibit different NER prop-

Findings of the Association for Computational Linguistics: NAACL 2025, pages 6822–6843

erties depending on its contextual usage. Consider the following examples: "[বাংলাদেশিরা]PER অতিথিপরায়ণ।" (trans:"The Bangladeshis are hospitable." [1]) and "আমি একজন [বাংলাদেশি]GPE।" (trans:"I am a Bangladeshi."). In the first case, "বাংলাদেশিরা" (trans:Bangladeshis; People living in Bangladesh are called Bangladeshis) refers to people (PER), while in the second, "বাংলাদেশি" (trans:Bangladeshi; Nationality of Bangladesh) denotes a geopolitical entity (GPE). To tackle these complexities, quality data annotation requires contextual data annotation and validation with a trained group of people who has core knowledge of language or domain. However existing Bangla NER datasets lack these considerations during their annotation process. To address this, we introduce **BanNERD**, the largest manually annotated and validated **Ban**gla **N**amed **E**ntity **R**ecognition **D**ataset to date which is constructed from a diverse array of sources across 29 distinct domains ensuring generalizability, comprehensiveness, and long-term applicability for various NER tasks. The dataset underwent meticulous human annotation and validation by a team of trained annotators and validators, adhering to a well-defined annotation guideline specifically designed for Bangla NER to tackle the aforementioned unique complexities. The annotation process is quality-controlled through an annotation management system tailored for Natural Language Processing (NLP) tasks. We employed an iterative annotation approach, as outlined in Islam et al. (2023), which progressively enhanced annotation consistency and quality, resulting in a commendable Inter-Annotator Agreement (IAA) score of 0.88. This will ensure that future researchers can focus on developing better Bangla NER systems instead of dealing with the time-consuming and expensive data annotation process.

In addition to a gold-standard Bangla NER dataset, we aimed to create a state-of-the-art lightweight NER recognition system for Bangla language deployable in small consumer hardware to facilitate real-world usage. To this endeavour, we introduce **BanNERCEM** (Bangla NER context-ensemble Method), which leverages contextual information from an external knowledge base (KB), such as Wikipedia, to enhance entity recognition. Unlike current state-of-the-art approaches (Wang et al., 2021, 2022b), which concatenate all contextual data along with the input sentence before feeding it into the NER model, our method processes each context individually through the NER model. To determine the final classified entity span, we employ a majority voting mechanism inspired by Yamada et al. (2020). Our approach modifies this majority voting process by initially conducting voting among contexts that directly contain the named entity present in the input sentence. Subsequently, voting is performed among the remaining predicted spans from other contexts. We argue that as a maximum input length of LLMs can cut off external contexts, this hierarchical voting strategy ensures no such cut-off leading to more effective utilization of contextual information. Recent LLMs like LLAMA-3.1 (Dubey et al., 2024) do not have this context cut-off problem due to their massive max token length - but have billions of parameters requiring significantly large GPU memory and high inference time. Instead, our method uses pre-existing lightweight Bangla LLMs (Bhattacharjee et al., 2022; Conneau et al., 2020; Sarker, 2020) which have parameter counts only in millions, thus requiring significantly less GPU memory while achieving noteworthy performance.

We believe our two-fold contributions of creating a gold-standard dataset and developing a state-of-the-art Bangla NER approach will benefit future researchers by allowing them to focus solely on improving NER systems without the additional burden of dataset creation. Our main contributions in this study are outlined as follows:

- We introduce BanNERD, the largest manually annotated and validated NER dataset for Bangla, comprising 85,175 sentences (with an average of 11.55 words per sentence), labelled with ten distinct NER classes: Person, Organization, Geo-Political Entity, Location, Event, Number, Unit, Date & Time, Term & Title, Miscellaneous.
- A total of 51 rigorously trained annotators and 5 validators all with formal linguistic knowledge annotated and validated the dataset achieving a high Inter Annotator Agreement (IAA) score of 0.88 demonstrating superior annotation quality.
- Our *BanNERCEM* achieves a 90.49% macro f1 score on our BanNERD dataset and about

[1]We denote English translations with a "trans:" tag and International Phonetic Alphabet (IPA) as "ipa:" tag

3.24% more average macro f1 over other existing Bangla NER datasets than the previous SOTA NER approaches. We also achieved new state-of-the-art results on five out of six previous Bangla NER datasets. On an English dataset, our approach is competitive against others.

- For entities where information is present in the contexts, we achieve about 4% performance improvement than previous context-concatenation-based SOTA demonstrating better context utilization capability of our approach.

- We achieve the best cross-dataset performance where our *BanNERCEM* approach trained on BanNERD achieves about $\sim$ 5% higher average macro f1 than the second best one. It shows the generalizability and high annotation quality of BanNERD dataset.

2 Related Work

Early works in NER frequently leveraged the linear-chain Conditional Random Field (CRF) (Lafferty et al., 2001). Later works used transformer models to achieve the SOTA performance (Zhang et al., 2018). Zhou et al. (2022) proposed a noise-robust NER framework utilizing multiple base models and jointly optimizing those with task-specific loss. Recent state-of-the-art approaches like Wang et al. (2021) used Google search to retrieve and use external contexts for inputs and achieved SOTA performance. Later, Wang et al. (2022b) employed a Wikipedia KB for additional context with an iterative context retrieval strategy achieving SOTA performance. Tan et al. (2023) incorporated an entity-centric Wikidata KB and designed three different retrieval strategies to enhance context retrieval.

One of the first Bangla NER datasets was WikiANN (Pan et al., 2017) which consisted of 12,000 Bangla sentences. They translate mentions from other languages to English via a translator and then link them to an English Knowledge base to find the entity type. Naamapadam (Mhaske et al., 2023) is another Bangla dataset with almost a million sentences. But, their training set is annotated via a pre-trained English NER tagger [2] and the test set though human annotated, only contains 607 sentences. MultiCoNER datasets (Malmasi et al., 2022a; Fetahu et al., 2023a) from the SemEval 2022 and 2023 competitions (Malmasi et al., 2022b; Fetahu et al., 2023b) follow the same approach of WikiANN of getting the entity mentions from an external knowledge base like Wikipedia. Karim et al. (2019) created the first human-annotated Bangla consisting of nearly 70,000 sentences with four entity types. Haque et al. (2023) created another human-annotated Bangla dataset consisting of 22,144 manually annotated Bangla sentences that are categorized into eight different entity types.

All the aforementioned datasets lack consideration for Bangla's unique linguistic properties Haque et al. (2023). As mentioned in **Section 1**, there is a high degree of inflections in the Bangla language which change the entity type of the same word based on the context. Haque et al. (2023) employed humans to annotate the dataset but only a small portion was validated by a linguist. Their annotation guideline does not specify how to tackle unique challenges faced in the Bangla language like an abundance of polysemous and homonymous words (Karim et al., 2019), inflections in proper nouns (Afrin et al., 2023) and multi-word expressions. Karim et al. (2019) discusses the challenges of annotating NER in the Bangla language but does no tackle them. They also do not validate their annotations by a validator. The MultiCoNER and other automatically annotated datasets use English gazetteer and some sort of alignment or mapping to English entities which is noisy in itself. In short, existing Bangla NER datasets are either noisy or do not consider unique challenges in annotating NER for the Bangla language.

BanNERD is the first dataset that takes into account all these challenges. Our annotation guideline is prepared by an expert team of linguists carefully considering all complexities and edge cases of Bangla NER annotation. The whole annotation is carried out by an expert team of annotators and validators maintaining the highest possible quality. BanNERD is also the largest among the two existing human-annotated datasets (Haque et al., 2023; Karim et al., 2019).

3 Development of BanNERD

3.1 Source selection, data collection, pre-processing, and data selection

We collected unprocessed raw textual data from popular websites containing Bangla texts. We se-

[2] https://huggingface.co/dslim/bert-base-NER

lected 29 sources covering 29 different domains provided in detail in **Appendix A**. We explain in detail how we avoid ethical and copyright issues during our data collection in **Section 3.2** and in a separate **Ethical Consideration** section.

The unprocessed raw data was processed following several steps including i) Unicode normalization, ii) HTML tags and URL removal, iii) language detection using Langdetect[3] and filtering out non-Bangla texts, iv) redundant whitespace removal, v) redundant punctuation mark removal, and vi) text deduplication using Jaccard similarity with a threshold of 0.95. We split the processed text into sentences and pre-tokenize the sentences to select as samples. These preprocessing steps are described in detail at **Appendix B**.

At first, we randomly selected about 15% sentences of total targeted data. We manually inspected the annotation of these selected data and found that many sentences contained no named entity. Hence, we employed a NER model following the work of Zhou and Chen (2021) in the data selection pipeline to filter out sentences containing no named entity. However, we randomly retained about 10% of total sentences with no named entity. More details on this filtering process is given in **Appendix C**.

3.2 Data Anonymization

We followed the anonymization process of (Volodina et al., 2020) and tailored it for our specific task. For sensitive personally identifiable information (PII) info like bank account, license numbers, etc we used Microsoft's Presidio tool [4] and also instructed human annotators to discard sentences containing any PII. The complete anonymization process is given in **Appendix D**.

3.3 Annotation guideline

Before commencing the annotation process it is essential to have proper annotation guidelines. So we adopted a comprehensive annotation guideline [5] which was developed by linguists and was validated and accepted by a national linguists and technical expert committee and then shared with the annotators. A summarized version of the guideline is provided in **Appendix F** with a description of each entity type. We have incorporated 10 named entity classes used in the Bangla language including 1. Person (*PER*), 2. Organization (*ORG*), 3. Geo-Political Entity (*GPE*), 4. Location (*LOC*), 5. Event (*EVENT*), 6. Number (*NUM*), 7. Unit (*UNIT*), 8. Date & Time (*D&T*), 9. Term & Title (*T&T*), and 10. Misc (*MISC*).

3.4 Annotation and Validation

We use the annotation platform used by Islam et al. (2023) as it had a superior performance tracking of annotators, IAA calculation compared with other annotation platforms. The dataset underwent annotation by a team of 5 validators possessing a master's degree in linguistics and 51 trained annotators, each possessing a minimum formal education. Other information regarding annotators and validators is available at **Appendix E**.

Annotations are carried out in groups of three annotators, each supported by a validator following the iterative annotation approach of Islam et al. (2023). For each iteration, a randomly selected chunk of 500 sentences is assigned to the group. Each sentence is independently annotated by all three annotators following the annotation guidelines. The majority voting among these 3 annotations determines the preliminary label for each instance. Subsequently, the assigned validator for that group reviews the annotations, corrects any errors, and finalises the labels. The annotation platform shows the accuracy of annotators. If any annotators require further training or the validator deems any annotations to require re-evaluation or identifies the need for additional training within the group, a collaborative session is organized. During these sessions, the group discusses the identified issues and addresses any complexities or misunderstandings. Some of these complexities are given in **Appendix G**. This iterative process is repeated throughout the annotation phase to ensure high-quality and consistent annotations.

Every validator undergoes training and supervision by linguists having a master's degree and prior experience in the field of linguistics. We adopt the BIO tagging scheme as it is widely used for postpositional languages including Bangla, Hindi, etc (Haque et al., 2023). This BIO tagging scheme is automatically converted from the annotation platform. We are following span-based NER annotation rather than character-based annotation.

[3] https://github.com/Mimino666/langdetect
[4] https://microsoft.github.io/presidio/
[5] BanNERD Bangla NER Annotation Guideline

Class	# Tokens	Percentage	IAA
PER	78,657	33.89	0.96
NUM	36,776	15.85	0.906
ORG	32,706	14.09	0.85
D&T	22,930	9.88	0.887
LOC	13,707	5.91	0.835
GPE	13,338	5.75	0.811
EVENT	13,203	5.69	0.824
UNIT	8,552	3.68	0.873
MISC	7,165	3.09	0.856
T&T	5,059	2.18	0.755

Table 1: Class distribution of BanNERD and class-wise IAA score. The 'others' tag is excluded as it is not a named entity. Percentages are calculated among named entity tokens only.

3.5 Statistical Analysis

Human annotators manually labeled $991,441$ tokens among which $232,093$ tokens are named entity tokens. The annotations are validated by linguistic experts, resulting in a dataset comprising $85,175$ sentences with an average sentence length of 11.55 words. The inter-annotator agreement (IAA), assessed using Cohen's kappa score was 0.88. The NE class distribution and class-wise IAA score are shown in **Table 1**. We found six other NER datasets in the Bangla language which we compare with BanNERD in **Table 2**. Haque et al. (2023), Karim et al. (2019) and BanNERD are the only human-annotated datasets in Table 2. BanNERD is the biggest human-annotated Bangla NER dataset. It also contains the highest variety of entity types excluding Fetahu et al. (2023a) which as mentioned previously is automatically annotated using English Wikipedia and translation to map English entities to Bangla entities.

4 Methodology

In this section, we present the functioning of our *BanNERCEM* system. Our motivation for creating BanNERCEM was to create a consistent system that achieves superior performance regardless of dataset and entity types while keeping the whole system lightweight which can be used in smaller consumer hardware. **Table 3 and 5** shows that our approach achieves both aforementioned objectives. We provide the architecture of our approach in **Figure 1**. The process begins with an input sentence comprising N tokens, denoted as $x = \{x_1, x_2, \ldots, x_n\}$. For each sentence, we extract top-K sentences related to the input from a knowledge base (KB) as the external context from the *Context Retrieval Module (CRM)*. Each extracted context is concatenated to the input sentence separately and then passed to the *Named Entity Recognition (NER)* module to identify the named entities. The CRM module uses WikiDump from Wikimedia[6] as a knowledge base (KB). We use the April 1st, 2024 version of WikiDump, convert it to plain text, and preprocess it. It is similar to Wang et al. (2022a) and explained in depth at **Appendix H**.

The NER module comprises of a Pretrained Language Model (PLM) followed by a CRF layer acting as the NER token classification head. Specific choice of PLM is given in **Section 5.3**. As said earlier, we pass each context from CRM module separately to the backbone and get K predictions in total from K contexts. A two-stage majority voting is conducted among these K predictions. In the first stage, the predicted entities are checked against their respective context to determine whether the entity **appears** in the context. Predicted spans from such contexts are accumulated and the majority voting among them defines the final entity. In the second stage, a majority voting among remaining spans where the predicted entity span **does not appear** in the context is conducted. The reasoning behind this hierarchical two-stage approach is that the CRM module is not error-proof and can contain contexts structurally similar to the input sentence but does not provide any extra information about the entities (Wang et al., 2022a). The first stage circumvents this issue by prioritizing predictions from contexts that contain information regarding entities.

5 Experiments

5.1 Datasets

To show the effectiveness and generalizability of our approach, we experiment on our BanNERD dataset along with the only available four other previous Bangla NER datasets Karim et al. (2019), B-NER (Haque et al., 2023), MultiCoNER I (2022) (Malmasi et al., 2022a), MultiCoNER II (2023) (Fetahu et al., 2023a), Naamapadam (Mhaske et al., 2023) and WikiANN (Pan et al., 2017). Data statistics of each dataset are provided in **Table 2**.

5.2 Baselines

We compare the *BanNERCEM* model with previously tested approaches in Bangla language and state-of-the-art NER approaches in English

[6] https://dumps.wikimedia.org/

Attribute	BanNERD	Haque et al. (2023)	Karim et al. (2019)	Malmasi et al. (2022)*	Fetahu et al. (2023a)*	Mhaske et al. (2023)*	Pan et al. (2017)*
Sentences	**85,175**	22,144	71,284	149,219	30,074	967,145	12,000
Tokens	991,441	297,418	983,663	896,116	393,509	15,419,213	51,618
Unique Tokens	73,526	34,237	96,155	83,986	42,619	525,515	8,191
Sentence Length	[3-20]	[1-233]	[5-30]	[1-35]	[2-85]	[1-100]	[3-62]
Avg. Sentence Length	11.55	13.43	13.80	6.005	13.08	15.94	4.30
Entities	**10**	8	4	6	34	3	3
Tagging Scheme	BIO	BIO	BIO	BIO	BIO	BIO	BIO
Number of Tags	21	17	9	13	67	7	7

Table 2: Comparison of BanNERD and existing Bangla NER datasets. * marked ones are automatically annotated via Knowledge base, translation or other methods. Among human annotated datasets, BanNERD is the largest and has the highest entity variety with 10 unique entity types.

NER. Approaches used in Bangla language but not limited to are - BiLSTM-CRF, BERT-CRF, BERT-BiLSTM, Banner (Ashrafi et al., 2020) and RaNER (Wang et al., 2022a) (a context-based NER approach that retrieves the top 10 contexts from Wikipedia Dump and concatenates it to the input sentence). English state-of-the-art approaches are - Noisy Label (Zhou and Chen, 2021), DiffusionNER (Shen et al., 2023) and Binder (Zhang et al., 2023). We also compared these approaches with GPT-4o (OpenAI, 2024a) and GPT-3.5 Turbo (OpenAI, 2024b) from OpenAI. We have followed the prompting strategy of (Lai et al., 2023) with ten few-shot examples which have been shown to be an effective prompting strategy in multilingual settings. Complete details of each baseline and the prompt construction process for OpenAI models can be found in **Appendix I**. We were unable to reproduce some state-of-the-art English methods particularly Luke (Yamada et al., 2020), due to specific entity-aware pre-taining done on English data which are unavailable in Bangla and too computationally expensive to reproduce. U-RaNER (Tan et al., 2023), a state-of-the-art method on Fetahu et al. (2023a) dataset did not open-source their whole approach and we excluded it from our experiments as we could not faithfully reproduce it. Other recent English NER approaches were created for few-prompt settings which are excluded from our setup as they are shown to underperform against finetuned models in their studies (Wang et al., 2023).

5.3 Training Strategy

For each respective baseline, we use the hyperparameters specified in their respective works without any hyperparameter tuning. We provide a brief description of their hyperparameter settings in **Appendix J** and refer to their works for detailed hyperparameter settings.

As different datasets have achieved the best performance using different PLMs, we use the respective PLM used for each dataset for a fair comparison with our approach. We use the BanglaBERT (Bhattacharjee et al., 2022) model for BanNERD, Karim et al. (2019) and WikiANN (Pan et al., 2017) dataset. For MultiCoNER datasets (Malmasi et al., 2022b; Fetahu et al., 2023a), we use the XLMRoBERTa (XLM-R) (Conneau et al., 2020) model as used in RaNER model (Wang et al., 2022a). For Haque et al. (2023) achieved highest performance using IndicbnBERT[7]. But as the resulting transformer model is now unavailable on Huggingface[8], we have used their second best-performing transformer model, SagorBERT (Sarker, 2020) for our experiments. For Naamapadam (Mhaske et al., 2023) dataset, we use bert-base-multilingual-cased [9] model.

For our *BanNERCEM* approach, we set a learning rate of $5e{-}6$ for the PLM model and $5e{-}2$ for the CRF layer following Wang et al. (2021) and a linear warmup-decay learning scheduler. We choose AdamW optimizer (Loshchilov and Hutter, 2019), train for 40 epochs and choose the best

[7] https://huggingface.co/neuralspace-reverie/indic-transformers-bn-bert

[8] https://huggingface.co

[9] https://huggingface.co/google-bert/bert-base-multilingual-cased

Method	BanNERD[1]	Haque et al. (2023)[2]	Karim et al. (2019)[1]	Malmasi et al. (2022)[3]	Fetahu et al. (2023a)[3]	Mhaske et al. (2023)[4]	Pan et al. (2017)[1]	AVG.
GPT-3.5 Turbo	40.94	41.51	40.94	31.85	9.76	36.72	64.51	37.12
GPT-4o	56.91	53.60	49.57	20.59	39.37	53.27	78.21	**50.21**
DiffusionNER	84.05	65.46	69.05	25.84	32.70	74.04	92.30	63.34
BiLSTM-CRF	74.31	58.00	62.28	57.51	61.39	73.18	91.17	68.26
BINDER	82.41	82.50	68.40	57.92	63.66	74.20	95.85	74.99
BERT-CRF	81.07	77.00	62.79	69.88	66.46	82.11	97.17	76.64
Noisy	88.78	79.80	67.53	72.09	56.41	82.50	97.11	77.74
BERT-BiLSTM	87.36	77.00	65.96	71.95	66.46	79.48	97.32	77.93
Banner	87.36	72.03	65.96	71.95	72.21	80.06	96.69	78.03
RaNER	$88.00^{\triangle+6.93}$	$78.42^{\triangle+1.42}$	$67.67^{\triangle+4.88}$	$83.51^{\triangle+13.63}$	$72.75^{\triangle+6.29}$	$82.61^{\triangle+0.5}$	$93.91^{\triangle-3.26}$	$80.98^{\triangle+4.34}$
BanNERCEM-NEA	$89.43^{\triangle+8.36}$	$82.82^{\triangle+5.82}$	$70.19^{\triangle+7.40}$	$82.25^{\triangle+12.37}$	$77.93^{\triangle+11.47}$	$82.52^{\triangle+0.41}$	$98.25^{\triangle+1.08}$	$83.34^{\triangle+6.7}$
BanNERCEM	$\mathbf{90.49}^{\triangle+9.42}$	$\mathbf{83.62}^{\triangle+6.62}$	$\mathbf{72.90}^{\triangle+10.11}$	$\mathbf{84.17}^{\triangle+14.29}$	$\mathbf{78.08}^{\triangle+11.62}$	$\mathbf{84.87}^{\triangle+1.7}$	$\mathbf{98.75}^{\triangle+1.58}$	$\mathbf{84.12}^{\triangle+7.48}$

Table 3: Entity-label macro F1 scores averaged across 4 runs with different random seeds for all methods on recent Bangla NER datasets. $\triangle$ denotes the difference of macro F1 compared with the baseline BERT-CRF model. 1 denotes BanglaBERT (Bhattacharjee et al., 2022) model, 2 denotes bangla-bert-base (Sarker, 2020) model, 3 denotes XLMRoBERTa (XLM-R) (Conneau et al., 2020), 4 denotes bert-base-multilingual-cased model (Devlin et al., 2019) is used. In *BanNERCEM*-NEA, **NEA** denotes *BanNERCEM* method without our entity aware majority voting using traditional majority voting (Yamada et al., 2020) and *BanNERCEM* denotes our complete approach with entity aware majority voting system.

model based on validation macro f1 scores. We split BanNERD in *85:10:5* ratio for the train, test, and validation splits respectively using stratified splitting. We train all models over 4 random seeds and average their results. We use a single NVIDIA A40 48GB GPU for all training and inference tests. More details regarding the experiment settings can be found in **Appendix J**.

6 Results and Analysis

6.1 Performance

We show the results of all approaches in **Table 3** for the Bangla NER datasets mentioned in **Section 5.1**. Performance is calculated based on entity-wise macro f1 scores as used in previous NER evaluations (Wang et al., 2022a; Tan et al., 2023) for all our evaluations. **Table 3** shows two variants of our approach. **BanNERCEM** denotes our full approach with hierarchical majority voting. **BanNERCEM-NEA** uses traditional majority voting (Yamada et al., 2020) instead of the hierarchical one proposed by us. **BanNERCEM** achieves SOTA results on our BanNERD dataset and five other previously published datasets (Karim et al., 2019; Haque et al., 2023; Malmasi et al., 2022b; Mhaske et al., 2023; Pan et al., 2017) achieving the highest average performance achieving 84.12% macro f1 score which is 3.14% greater than the similar context based RaNER method and 7.48% greater than the BERT-CRF model. **Table 3** also shows using our hierarchical voting approach increases average macro f1 score by 0.78% compared with convention majority span voting of Yamada et al. (2020). On the WikiANN (Pan et al., 2017) dataset, all the models' performances are similar. We rechecked their test dataset and found that most of the sentences are only common entity names instead of a full sentence which may have led to a high performance across the board.

On the English portion of the MultiCoNER I dataset (Malmasi et al., 2022b), we achieved a macro F1 score of 90.34%, closely approaching the state-of-the-art (SOTA) result of 91.21% achieved by the RaNER model. However, as our primary focus was developing a robust NER system for Bangla, effectiveness with English and other languages is left for future work.

6.2 Analysis

Performance of Other NER Approaches in Bangla Datasets The RaNER model (Wang et al., 2022a) performs slightly worse on our BanNERD, B-NER dataset (Haque et al., 2023) and WikiANN dataset (Pan et al., 2017) than the Noisy Label Model (Zhou and Chen, 2021) although the latter is not using any external context. The Noisy Label model (Zhou and Chen, 2021) performs poorly on the MultiCoNER 2023 dataset (Fetahu et al., 2023a) although performing competitively on other datasets.

DiffusionNER and BINDER demonstrated inferior performance compared to simpler BERT-CRF architecture. For DiffusionNER, this can be associated with slow convergence dynamics as increas-

Dataset	RaNER	BanNERCEM	BanNERCEM-NEA
BanNERD	95.27	**95.31**	94.88
Haque et al. (2023)	83.48	**89.72**	89.32
Karim et al. (2019)	78.51	**82.78**	82.24
Malmasi et al. (2022b)	88.60	**89.86**	85.95
Fetahu et al. (2023a)	77.74	**85.12**	84.15
AVG.	84.72	**88.56**	87.30

Table 4: Comparison between RaNER (Wang et al., 2022a) and our *BanNERCEM* approach on those entities that the retrieved contexts have information about. Our approach performs considerably better than RaNER on all datasets.

LLM	Sequence Length	FLOPS (in GFLOPS)	MACs (in GMACS)	GPU Memory (in MB)
LLAMA-3.1 (8.03B)	128	1,920	960.6	32,630.63
	256	3,840	1,920	34,260.63
	512	7,690	3,840	38,032.63
	1024	15,370	7680	44,080.63
	2048	-	-	>48,000
BanglaBERT (110.03M)	128	22.36	11.17	842.63
	256	45.94	22.95	1,012.63
	512	96.72	48.32	1,210.63
XLM-RoBERTa-large (559.89M)	128	78.96	39.46	2,688.63
	256	161.15	80.53	2,962.63
	512	335.24	167.5	3,640.63

Table 5: Comparing LLAMA-3.1 8B model with pretrained Bangla LLMs like BanglaBERT and XLM-RoBERTa-large. LLAMA-3.1 with 2,048 tokens results are empty due to cuda memory overflow issue. Parmaeters count are given beside LLM name in brackets. (B) dentoes a billion and (M) denotes a million.

ing the number of denoising timesteps from the default 1,000 (as proposed in the original Diffusion-NER framework) to 2,000, we observed a notable gain of 8.93% in average macro F1 scores. For BINDER, we added a single example to each entity type description and found significant improvements in macro F1 scores, with gains of 1.59%, 3.61%, and 0.93%, respectively in our BanNERD and MultiCoNER (Malmasi et al., 2022a; Fetahu et al., 2023a) datasets. Further experimentation settings are out of scope for our research.

For lightweight networks like BiLSTM, the performance is very poor compared with the transformer-based methods. The results from GPT-4o and GPT-3.5 are poorer than the fine-tuned results in **Table 3**. These results are comparable to the results reported by Tan et al. (2023) on MultiCoNER II (Fetahu et al., 2023a) dataset where they reported a macro F1 score of 14.76%. We test GPT-4o with zero-shot setting on our BanNERD dataset where the macro f1 score decreased from 56.91% to 49.51%. We concur that instruction tuning on the annotated NER labels is required for these LLMs to perform comparatively in Bangla NER classification.

On MultiCoNER 2023 dataset (Fetahu et al., 2023a), we were unable to reproduce the current SOTA method U-RaNER's (Tan et al., 2023) results because their retrieval system was not made publicly available. Using our retrieval approach, U-RaNER achieves 72.43% macro f1 score on the MultiCoNER 2023 dataset (Fetahu et al., 2023a) which is less than their reported result of 81.60%.

Better Context Utilization We further analyze the context utilization capability of our BanNERCEM approach by comparing its performance on those entities whose information has been found in the retrieved contexts. **Table 4** shows that **BanNERCEM** outperforms **RaNER** on all datasets in this experiment achieving $\sim$ 4% greater average macro f1 score. The average performance on entities that the contexts have information about is 6.71% higher than the overall performance reported in **Table 3**. As these entities are present in the context, our entity-aware majority voting will prioritize the predictions from these contexts. **Table 4** shows that this approach increases the performance on these entities by 1.26% average macro f1-score. On MultiCoNER I (Malmasi et al., 2022b) dataset, **BanNERCEM** approach gives 3.91% macro f1 score gain over **BanNERCEM-NEA**. This shows that context ensembling and entity-aware voting combinedly contribute to better context utilization leading to greater performance.

Using Current LLMs As current LLMs have a massive max input length, they can utilize all external contexts from KB without breaking them into separate parts like our approach. But recent studies on the Bangla language (Mahfuz et al., 2024) suggest that small pretrained models when finetuned, perform competitively with models like LLAMA-3.1 8B and 70B models (Dubey et al., 2024). We provide a comparison of floating-point operations per second (FLOPS), multiply-accumulate operations (MACs) count and GPU memory usage in **Table 5**. Our end goal is to create a lightweight, high performant NER model on Bangla runnable on smaller consumer hardware which will be difficult with LLAMA like LLMs. Hence, we did not fine-tune LLAMA-like models with and without retrieved contexts.

Trained on Datasets	Evaluated on Datasets					
	Karim et al. (2019)	Haque et al. (2023)	Malmasi et al. (2022)	Fetahu et al. (2023a)	Ban NERD	AVG.
Karim	**72.90**	64.81	47.17	51.43	57.74	55.29
Haque	54.40	**83.62**	41.34	49.89	32.49	44.53
Malmasi	38.73	49.34	**84.17**	83.43	43.05	53.54
Fetahu	41.85	57.81	78.52	**78.08**	46.73	56.23
BanNERD	67.63	61.24	52.99	65.75	**90.49**	**61.90**

Table 6: Cross Dataset result using our **BanNERCEM** approach trained on BanNERD. The AVG. column contains the average of macro F1 scores excluding the diagonal values as the diagonal values contain macro F1 scores on their own dataset. Our approach trained on BanNERD dataset achieves the highest average macro F1 score among all other datasets.

6.3 Cross Dataset Evaluation

We perform a cross-dataset experiment using our **BanNERCEM** approach to measure dataset generalizability. If a dataset is generalized, a model trained on that dataset should achieve greater performance on other NER datasets than models trained on a different dataset without any further fine-tuning. The Bangla NER datasets we have experimented on have three common NER labels: Location, Person and Organization/Corporation on which we perform our cross-dataset evaluation. **Table 6** contains the top 5 result of this cross-dataset evaluation experiment. We see that the BanNERD trained **BanNERCEM** model achieves the highest average macro F1 score among all datasets. BanNERD achieves a 5.67% greater average macro F1 score than the second-placing MultiCoNER II (Fetahu et al., 2023a) dataset. This shows that the BanNERD-trained model has achieved superior generalizability. Our BanNERD dataset is carefully annotated with multiple human annotators and linguists which has contributed to high cross-dataset performance.

6.4 Ablation Study

Impact of Number of Retrieved Contexts We analyze the performance of our approach for different values of $K \in 3, 6, 10$ in the top-K retrieval of contexts. **Table 7** shows that a higher number of contexts results in performance gains on all datasets. The average performance gain over all datasets from using 3 contexts to 10 contexts is 2.14% and 6 to 10 is 1.28%. **Table 7** also shows the entity coverage of using various number of contexts. Entity coverage is the percentage of the total number of gold entities found in the retrieved contexts. Higher entity coverage means the retrieved contexts are more correlated with the input sentence as they contain more information regarding the entities. **Table 7** shows that increasing context number from 3 to 10 results in an average gain of $\sim$ 10% in entity coverage. It also shows that using only 3 contexts, our approach outperforms RaNER on all datasets except MultiCoNER I (Malmasi et al., 2022b). This shows that our *BanNERCEM* approach is utilizing the retrieved contexts better than RaNER. We did not perform this experiment on Naamapadam (Mhaske et al., 2023) and WikiANN (Pan et al., 2017) datasets due to their smaller test set size but large training computation requirements.

Dataset	Macro F1			Entity Coverage(%)		
	Size-3	Size-6	Size-10	Size-3	Size-6	Size-10
BanNERD	89.92	90.32	**90.49**	41.14	47.91	53.48
Haque et al. (2023)	82.38	82.51	**83.62**	31.13	39.46	46.78
Karim et al. (2019)	70.45	71.41	**72.90**	57.75	62.94	66.26
Malmasi et al. (2022b)	82.77	83.47	**84.17**	42.84	51.64	54.99
Fetahu et al. (2023a)	77.87	77.89	**78.08**	55.07	58.72	59.99
AVG.	67.74	68.60	**69.88**	50.67	56.53	60.63

Table 7: Performance of our *BanNERCEM* approach with top-3,6 and 10 contexts. Increasing context size increases entity coverage and increases the performance in all datasets.

7 Conclusion

In this paper, we provide a gold standard Bangla Named Entity Recognition dataset with 85,175 sentences and ten types of entities. We also propose a context-based approach for Named Entity Recognition where we passed each context separately and used an entity-aware majority voting among predicted spans to calculate the final predictions. We achieve competitive performance when compared with the contemporary context-based approaches which concatenate the contexts directly with the input. On manually human-annotated Bangla datasets, our method outperforms all other approaches. We also show that our proposed approach can utilize the contexts better than conventional context concatenation approaches. Our current retrieval system only uses Bangla Wikipedia Dump as the knowledge base. We plan to enhance our retrieval system to incorporate more knowledge bases. We also plan to explore more efficient context utilization methods to reduce our training and inference time.

Limitations

On MultiCoNER datasets, performance on the 'Creative Work (CW)' labels is the lowest. These CW tags are names of books, movies, arts, etc. which may not look like traditional named entities (Malmasi et al., 2022b). For example, in the sentence 'ঝুলন্ত স্ক্রোল ল্যান্ডস্কেপ কাকেজিকু কাগজে কালি এবং হালকা রঙ' (ipa:ɟʱulonṭo skrol lændʃkep kakeɟiku kagoɟe kali ebɔŋ halka rɔŋ) (trans: Hanging scroll landscape ink and light color on kakejiku paper), 'কাকেজিকু' (ipa:kakeɟiku) (trans:kakejiku) is a ArtWork and belongs to CW tag. This word is fairly uncommon which our approach mistakenly predicts as 'WrittenWork'. On these CW tags, we achieve 76.13% and 76.27% macro f1 scores on MultiCoNER I and II datasets respectively.

As we pass each context separately to the model for every input sentence, this increases the training and inference time of our approach. **Table 8** contains the speed of our *CRM* module and training speeds of RaNER and our *BanNERCEM* approach with top-3 and top-10 contexts. Training time for top-3 contexts is similar in RaNER and our approach. When top-10 contexts are used, RaNER takes 644 seconds less time per epoch. Previously, we have shown that using top-3 contexts, our approach has outperformed **RaNER** on all Bangla NER datasets except one. In order to expand our approach to a higher number of contexts, a more efficient method of combining contexts needs to be explored.

Module	Sentences/Second	Epoch Time
Context Retrieval from English KB	64.68	N/A
Context Retrieval from Bangla KB	222.24	N/A
RaNER@3 (training)	70.01	1,079
RaNER (training)	34.4	2,143
CE@3(training)	210.52	1,026
CE@10 (training)	258.34	2,787
BERT-CRF(training)	531.87	135

Table 8: Speed of the Context Retrieval Module and the context-based models. BERT-CRF is the baseline model which does not use an external context.

Ethical Considerations

We adhere to stringent ethical standards, aligning with the fair data usage policies of major social media platforms such as Facebook and YouTube. This commitment stems from the utilization of public funds provided by a national research funding authority. This body explicitly restricts commercial usage, with the dataset being exclusively allocated for non-commercial and academic research. To comply with Facebook[10] and YouTube's[11] fair usage policies, standard API calls were employed to collect bulk data from "Facebook public pages" and public YouTube comments. Subsequently, sentences were curated through random sampling, incorporating anonymization and de-identification processes. Notably, sensitive data was not obtained from any confidential social media pages. The funding body's regulations mandated the meticulous removal of personal mentions, including emails and phone numbers, through manual linguistic inspection before utilizing the BanNERD for any experiments in this paper. Access requests to the dataset, code and other materials will be entertained through an End-User License Agreement (EULA), subject to specific conditions. These conditions include the dataset's exclusive use for non-commercial research purposes, prohibition of redistribution by the dataset usage-grantee, and permission for illustrative purposes in publications for scientific and educational use. The research funding authority will impose additional conditions, such as a one-year validity period for the EULA, necessitating renewal of the data-sharing policy changes. Researchers are required to uphold the highest ethical standards in their research and publications, giving proper acknowledgment to the research funding body.

Acknowledgments

We would like to express our gratitude to the Bangla Syntactic Treebank Corpus With Processing Pipeline and Distribution Platform project, which is part of the Enhancement of Bangla Language in ICT through Research & Development (EBLICT) [12] initiative, supported by the Bangladesh Computer Council [13] under the ICT Division of the Government of Bangladesh [14]. We extend our sincere appreciation to the dedicated development consultant team at Giga Tech Limited [15] and the Dhaka University Linguistics and Communication Department [16] for their annotation support. Furthermore, we would like to acknowledge the invaluable financial support provided by the

[10] https://www.facebook.com/help/1020633957973118
[11] https://support.google.com/youtube/answer/9783148
[12] http://eblict.gov.bd/
[13] https://bcc.gov.bd/
[14] https://ictd.gov.bd/
[15] https://gigatechltd.com/
[16] https://www.du.ac.bd/body/LIN

People's Republic of Bangladesh, which enabled the successful execution of this research.

References

Sadia Afrin, Md Shahad Mahmud Chowdhury, Md Islam, Faisal Khan, Labib Chowdhury, Md Mahtab, Nazifa Chowdhury, Massud Forkan, Neelima Kundu, Hakim Arif, et al. 2023. Banlemma: A word formation dependent rule and dictionary based bangla lemmatizer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3695–3710.

Imranul Ashrafi, Muntasir Mohammad, Arani Shawkat, Galib Nijhum, Redwanul Karim, Nabeel Mohammed, and Sifat Momen. 2020. Banner: A cost-sensitive contextualized model for bangla named entity recognition. *IEEE Access*, PP:1–1.

Iyadh Ben Cheikh Larbi, Aljoscha Burchardt, and Roland Roller. 2023. Clinical text anonymization, its influence on downstream NLP tasks and the risk of re-identification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 105–111, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Pramit Bhattacharyya, Joydeep Mondal, Subhadip Maji, and Arnab Bhattacharya. 2023. VACASPATI: A diverse corpus of Bangla literature. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1130, Nusa Dua, Bali. Association for Computational Linguistics.

N. Chinchor and P. Robinson. 1998. Appendix E: MUC-7 named entity task definition (version 3.5). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der

Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xi-

aofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. MultiCoNER v2: a large multilingual dataset for fine-grained and noisy named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2027–2051, Singapore. Association for Computational Linguistics.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023b. SemEval-2023 task 2: Fine-grained multilingual named entity recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2247–2265, Toronto, Canada. Association for Computational Linguistics.

Md. Zahidul Haque, Sakib Zaman, Jillur Rahman Saurav, Summit Haque, Md. Saiful Islam, and Mohammad Ruhul Amin. 2023. B-ner: A novel bangla named entity recognition dataset with largest entities and its baseline evaluation. *IEEE Access*, 11:45194–45205.

Md. Ekramul Islam, Labib Chowdhury, Faisal Ahamed Khan, Shazzad Hossain, Md Sourave Hossain, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. Sentigold: A large bangla gold standard multi-domain sentiment analysis dataset and its evaluation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4207–4218, New York, NY, USA. Association for Computing Machinery.

Redwanul Karim, M. A. Islam, Sazid Simanto, Saif Chowdhury, Kalyan Roy, Adnan Neon, Md Hasan, Adnan Firoze, and Mohammad Rahman. 2019. A step towards information extraction: Named entity recognition in bangla using deep learning. *Journal of Intelligent & Fuzzy Systems*, 37:1–13.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Tamzeed Mahfuz, Satak Kumar Dey, Ruwad Naswan, Hasnaen Adil, Khondker Salman Sayeed, and Haz Sameen Shahgir. 2024. Too late to train, too early to use? a study on necessity and viability of low-resource bengali llms. *Preprint*, arXiv:2407.00416.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 task 11: Multilingual complex named entity recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1412–1437, Seattle, United States. Association for Computational Linguistics.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2024a. Gpt-4o. `https://platform.openai.com`. Accessed: 2024-10-12.

OpenAI. 2024b. Gpt-4o. `https://platform.openai.com/docs/models/gpt-3-5-turbo`. Accessed: 2024-10-12.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. DiffusionNER: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3875–3890, Toronto, Canada. Association for Computational Linguistics.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, and Yong Jiang. 2023. DAMO-NLP at SemEval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2014–2028, Toronto, Canada. Association for Computational Linguistics.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 357–369, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022a. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, et al. 2022b. Damo-nlp at semeval-2022 task 11: A knowledge-based system for multilingual named entity recognition. *arXiv preprint arXiv:2203.00545*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023. Optimizing bi-encoder for named entity recognition via contrastive learning. *Preprint*, arXiv:2208.14565.

Wenxuan Zhou and Muhao Chen. 2021. Learning from noisy labels for entity-centric information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew Arnold, and Bing Xiang. 2022. Learning dialogue representations from consecutive utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 754–768, Seattle, United States. Association for Computational Linguistics.

Appendix

A Source selection and raw data collection

We collected unprocessed raw textual data from various websites containing Bangla texts. During the selection of the sources, we focused on covering several domains and use cases from the socioeconomic perspective of Bangladesh such as health, letters, notices, manuals, public opinions, politics, etc. BanNERD covers 29 different domains across 29 different sources. The domain distribution and source distribution are depicted in **Figure 2** and **Figure 3** respectively. We also list all the sources that we use to collect data below:

- independent24.com

- ebanglalibrary.com
- mzamin.com
- bn.wikipedia.org
- prothomalo.com
- nctb.com
- bdlaws.minlaw.gov.bd
- anannya.bd
- somewhereinblog.net
- amrabondhu.com
- pmo.gov.bd
- mole.gov.bd
- bnpbd.org
- bdun.org
- nimc.gov.bd
- bitac.gov.bd
- banglanotice.com
- bn.banglapedia.org
- rokomari.com
- pressinform.gov.bd
- facebook.com
- banglacyber.com
- hazabarolo.com
- golperjhuri.com
- bn.quora.com
- manuals.plus
- storymirror.com
- Sherajobs.com
- corporatenews.com.bd

After selecting the sources, we employed several custom crawlers to collect the raw texts. The crawlers were developed using the Scrapy[17] framework.

B Data pre-processing and Selection

At first, the unprocessed raw data was processed following several steps including i) Unicode normalization, ii) HTML tags and URL removal, iii) language detection using Langdetect[18] and filtering out non-Bangla texts, iv) redundant whitespace removal, v) redundant punctuation mark removal,

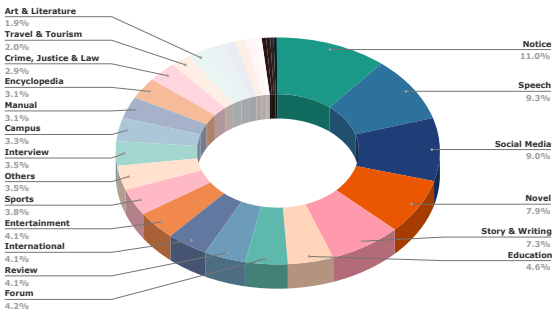

Figure 2: Distribution of the domains in BanNERD. The dataset covers 29 different domains where 'Notice & Circular' is the dominant one.

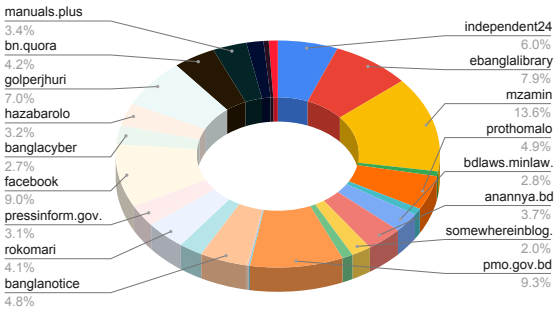

Figure 3: Illustration of our proposed dataset's source distribution which shows a uniform distribution across various sources.

and vi) text deduplication using Jaccard similarity with a threshold of 0.95. After getting the processed raw dataset, we split the text into sentences using regular expressions[19]. We pre-tokenized the sentences and selected them as samples for annotation in an iterative approach. At first, we randomly selected about 15% sentences of total targeted data. After the annotation, we manually inspected the annotated dataset and found that many sentences did not contain any named entity. So, we employed an ML model in the data selection pipeline to filter out the sentences not containing any named entity. However, we randomly retained about 10% sentences with no named entity. The further chunks of sentences were selected for the annotation process. During the selection of each chunk, we took approaches to make the dataset as distributed as possible across different domains and sources.

[17] https://scrapy.org/

[18] https://github.com/Mimino666/langdetect

[19] We considered the following punctuation marks as the of a sentence: [. (dot), ? (question mark), ! (exclamation mark), । (danda)].

C Data Filtering

We use a base NER model to remove any sentences without entities and then retain 10% of sentences randomly. The base NER model consists of a transformer model followed by a CRF layer (BERT-CRF). To ensure that it is capable of identifying complex entities, it was trained to only whether a word is a named entity or not excluding the entity type. For this, we concatenated the training set of Bangla NER datasets reported in **Section 5.1**. The base model achieved 89.24% accuracy on the concatenated test sets of all datasets. We also measured the accuracy of this base NER model after finishing the annotation of the first 1,000 sentences for a sanity check. It achieved 93.05% accuracy on this test set. There were 68,465 sentences tagged as having no entities during our filtering process from where we retained 6,847 sentences for annotation through 10% random retainment. After finishing the whole annotation, the BanNERD dataset contains 6,163 sentences without any entities out of 85,175 sentences. So, the base model's accuracy of finding sentences without entities is 90.01% on the whole BanNERD dataset. This high accuracy of the base NER model ensures that it can capture complex entities ensuring BanNERD dataset is a difficult NER dataset with hard to detect complex entities.

D Data Anonymization and De-identification

We followed the anonymization process of (Volodina et al., 2020) and tailored it for our specific task. For sensitive personally identifiable information (PII) info like bank account, license numbers, etc we used Microsoft's Presidio tool [20]. Presidio checks for the existence of such information in the dataset using regex and discards sentences containing any detected PII. Since Presidio does not support Bangla, we custom-built regex patterns for the Bangla language. There can be PIIs that were not detected via this tool. So, annotators were instructed to discard sentences if they saw the sentences contained any of the above PIIs. During this process, we discarded 1,592 sentences. In total, we discarded 12 Bank Account Numbers, 1,533 License numbers, 25 Emails, 33 Zip codes, 56 URLs, 2 Phone Numbers, 8 National Identification/Social Securirty numbers.

[20] https://microsoft.github.io/presidio/

For personal names, geo data, age and institutions/organizations, we perform a random replacement operation after the annotation process. In this process - person, geo-data or institution-related information is replaced randomly by an entity of the same entity type. This ensures reidentification remains very difficult on our BanNERD dataset.

Although no anonymization process is entirely accurate, we want to emphasize that our anonymization process is more rigorous compared to previous studies like (Ben Cheikh Larbi et al., 2023; Lison et al., 2021) where they mostly used automatic anonymization. In contrast, we use human annotators along with the automatic anonymization process to ensure that no sensitive information or identification of personnel is possible via the BanNERD dataset. We strongly believe that this ensures our BanNERD dataset's safety of usage.

E Annotator Details

We have chosen annotators who have received formal education in linguistics. The validators all possess a master's degree in linguistics. As we have prepared a Bangla NER dataset, all annotators and validators were Bangladeshi. Out of 51 annotators, 23 of them were female and 28 of them were male. Out of 5 validators, 3 of them were male and 2 of them were female. The annotators are renumerated hourly based on the local labour law.

F Summarized Guideline for Bangla Named Entities Recognition

In Bangla, 10 types of Named Entities are used as listed below:

1. **Person:** This entity refers to individuals, groups of people, fictional characters, designations, individual common names, etc. Example: "আলালের ভাই দুলাল এসেছে।" (tras:Alal's brother Dulal has arrived) here, আলালের(Alaler) and দুলাল(trans:Dulal) will be tagged as Person. However, multi-word expressions are widely used in Bangla text, especially in idioms, that convey entirely different from their literal interpretation, such as "আদর দিয়ে ছেলেকে আলালের ঘরের দুলাল করে তুলেছ" (trans:You have spoiled the boy.) where আলালের(trans:Alaler) and trans:দুলাল(trans:Dulal) are not taggable.

2. **Organization:** These entities refer to formally established associations such as businesses, industries, corporations, agencies, government units, sports teams, formally organized music groups, etc. Example: "আমি জাতীয় সংসদ ভবনের সামনে দারিয়ে আছি।"(trans:I am standing in front of National Parliament) The context of this sentence might create confusion between organization and location tag. However, name of the organizations always will be considered as Organization tag.

3. **Geo-Political Entity (GPE):** GPE are composite entities that include a population, a government, a physical location, an official name, a nation (or province, state, county, city, etc.), nationality, etc. Example: "চোখে পড়ে আর্জেন্টিনার ফ্যানদের আনন্দ উল্লাস।" (trans:You can see the joy of Argentina fans.) here, আর্জেন্টিনার (trans:Argentina) will be marked as GPE.

4. **Location:** Location entities refer to geographic or astronomical locations that do not form a political body such as address, celestial entity, local towns, villages, facilities, sea, river bodies, etc. Example: "তিনি দুর্গাপুর যাবার গাড়ির ব্যাপারে খোঁজ নিল।" (trans:He inquired about the vehicle to go to Durgapur) here, দুর্গাপুর (trans:Durgapur) considered as Location entity. However, because of uncertainty in entity types or variation in existing entity may cause tag overlap. Here, "তিনি দুর্গাপুর উপজেলায় বাস করে।" (trans:He lives in Durgapur Upazila.), the word upazila means an administrative body so here দুর্গাপুর (trans:Durgapur) will be tagged as GPE to mitigate the overlapping problem.

5. **Event:** Events indicate a specific occasion which may be a noteworthy occurrence, a social, national, or international activity, a program of sports, etc. Example: "বিক্ষোভটি সমিতি ভবনে সামনে হয়।" (trans:The movement took place in front of the association building.) here, বিক্ষোভটি(bikk^{h}ob^{ɦ}oti) (trans:The movement) will be considered as Event tag.

6. **Number:** Number indicates ordinal and cardinal numbers, integers, etc. Example: একাত্তরটি মামলা নিয়ে আলাপ হয় সেখানে। (trans:Seventy-one cases were discussed there.) here,একাত্তর (ækaṯṯor) (trans:Seventy-one) referring to the Number entity. However, in the second text "পুরস্কার হাতে নিয়ে একাত্তরের সাথে আলাপে এ কথা বলেন তিনি। (trans:He said this with Ekattor after receiving the award.) the context referring একাত্তর (ækaṯṯor) (trans:Seventy-one) as an organization.

7. **Unit:** Units are the entities used to express and distinguish measurements of money, number, rate, age, etc. Example: "রাজধানীতে ২০০৫ সালে প্লাস্টিক ব্যবহার হতো ৯.২ কেজি" ।(trans:In 2005, 9.2 kg of plastic was used in the capital) here, কেজি (keʄi) (trans:kg) is Unit.

8. **Date & Time:** It represents exact dates and times including duration, days of the week, month, year, and time of the day. Example: "২০২১ সালের ১৮ই আগস্ট এই আসরের সময়সূচী প্রকাশ করা হয়।"(trans:The schedule of this event was released on 18 August 2021.) here ২০২১ সালের ১৮ই আগস্ট(trans:18 August 2021) will be tagged as Date&Time.

9. **Term & Title:** This represents the terms and titles of different entities. The term describes a thing or concept such as the name of a theory, process name, method, invention, policy, etc. The title of different things such as the name of a book, song, or artwork such as a film, drama, novel, short film, piece of art, sculpture, etc will be included in this tag. Example: "এ আলোচনার মূল বিষয় থিওরি অব রিলেটিভিটি।"(trans:The main topic of this discussion is Theory of Relativity) here, থিওরি অব রিলেটিভিটি (trans:Theory of Relativity) is Term & Title.

10. **Misc:** This includes any entity that does not fit into the above entities such as religions, languages, products, awards, games, calamities, animals, and so on. Example: "ভারত চাল রপ্তানিতে শীর্ষে।" (trans:India is the top exporter of rice..) here, চাল (trans:rice) will be tagged as miscellaneous for being object but existing polysemous and homonymous words in Bangla creates ambiguity while detecting NER category. Here,দারুন এক দাবার চালে হারালেন বিপক্ষকে। (trans:He defeated his opponent with a brilliant chess move.) in this text চাল (trans:move) means move which is not a taggable named entity.

Query	Retrieval Strategy	Retrieved Cotnexts	Matched Parts
এর ফলে আগামী বছর বেকারত্বের হার বৃদ্ধি এবং অর্থনৈতিক মন্দার আশঙ্কায় ইউরোপীয় ইউনিয়ন । (trans:As a result, the European Union fears an increase in the unemployment rate and an economic recession next year.). Entities: আগামী বছর -(trans:Next year), ইউরোপীয় ইউনিয়ন - (trans:the European Union)	Sentence Retrieval	1. অর্থনৈতিক মন্দার ফলে এপ্রিল 2008 এবং এপ্রিল 2009 এর মধ্যে বেকারত্বের হার 4.7 থেকে 6.3% বৃদ্ধি পেয়েছে। (trans:The economic downturn caused the unemployment rate to increase from 4.7 to 6.3% between April 2008 and April 2009.) 2. এর ফলে রপ্তানি বৃদ্ধি এবং বেশ কয়েকটি উৎপাদনের খাত সৃষ্টি হয়েছে। (trans:This has resulted in increased exports and the creation of several manufacturing sectors.) 3. এর ফলে দুর্নীতি বৃদ্ধি পায় ও অর্থনৈতিক বৃদ্ধির হার মন্থর হয়ে পড়ে। (trans:As a result, corruption increased and the rate of economic growth slowed down.)	1. অর্থনৈতিক মন্দার ফলে (trans:economic downturn), এবং (and), এর (of), বেকারত্বের (trans:unemployment), হার (ratio), বৃদ্ধি (increase) 2. এর ফলে (trans:because of), বৃদ্ধি (trans:increase), এবং (and) 3. অর্থনীতি (trans:economy), এবং (and), বেকারত্ব (trans:unemployment), বৃদ্ধি (trans:increase)
	Entity Retrieval	1. ইউরোপীয় ইউনিয়ন বা ইইউ ইউরোপ মহাদেশের অধিকাংশ দেশের একটি অর্থনৈতিক ও রাজনৈতিক জোট। (trans:The European Union or EU is an economic and political union of most of the countries of the European continent.) 2. ১ নভেম্বর ১৯৯৩ সালে এই চুক্তি কার্যকর হয় যার ফলে 'ইউরোপীয় ইউনিয়ন' এবং ইউরোপের একক মুদ্রা হিসেবে 'ইউরো' চালু হয়। (trans:This agreement entered into force on 1 November 1993, creating the 'European Union' and the 'Euro' as Europe's single currency.) 3. আগামী প্রকাশনী বাংলাদেশ জ্ঞান ও সৃজনশীল প্রকাশক সমিতির সদস্য। (trans:Next Publication is a member of Bangladesh Knowledge and Creative Publishers Association.)	1. ইউরোপীয় ইউনিয়ন (trans:European Union), ইউরোপ (trans:Europe), অর্থনৈতিক (trans:Economy) 2. যার ফলে ইউরোপীয় ইউনিয়ন এবং ইউরোপের (trans:creating the 'European Union' and the 'Euro') 3. আগামী (Tomorrow)

Table 9: Example of different retrieval strategies for the sentence "এর ফলে আগামী বছর বেকারত্বের হার বৃদ্ধি এবং অর্থনৈতিক মন্দার আশঙ্কায় ইউরোপীয় ইউনিয়ন (trans:As a result, the European Union fears an increase in the unemployment rate and an economic recession next year)" where আগামী বছর (Next year) and ইউরোপীয় ইউনিয়ন (trans:European Union) are entities belonging to 'Date and Time' and 'Organization' types respectively.

G Complexities in Bangla NER Annotation

1. **Person and Term & Title complexity:** Fictional characters from books, movies and TV shows are considered Person. For example, in the sentence: সুপারম্যান এর কাছে এটা কোনো ব্যাপার না (trans: It doesn't matter to Superman.) - সুপারম্যান (trans: Superman) is a fictional character and is considered "Person". But, this creates complexities for movie, book titles that sound like a fictional character name. For example, in this sentence: অপুর সংসার কে সত্যজিৎ রায়ের শ্রেষ্ঠ সৃষ্টি বলে বিবেচনা করা হয়। (trans: Apur Sansar is considered to be Satyajit Ray's finest work.), অপুর সংসার (trans: Apur Sansar) is a movie title and should be tagged under "Term & Title". But অপুর individually sounds like a person which often can create ambiguity. In these instances, annotators are encouraged to search the title name on the web if they are not sure if it is a title of a movie/book, etc.

2. **Organization and GPE complexity:** Sometimes, an organization can include names of countries. For example, ভারতীয় সেনাবাহিনী (trans: Indian Army) as a whole is the name of an Organization, so it is tagged as "Organization". But if we change this to ভারতের সেনাবাহিনী (India's Army) - then it is no longer a name of an organization as a whole. Hence, "India" will be separately tagged as "GPE" as it is the name of a country.

The exhaustive list of all complexities are provided in the complete Annotation guideline [21] with clear examples that make them understandable to the annotators.

H Context Retrieval Module (CRM)

The *CRM* module uses WikiDump from Wikimedia[22] as a knowledge base (KB). We use the April 1st, 2024 version of WikiDump, convert it to plain text, and preprocess it. Every Wikipedia document is defined by three key fields: sentence, paragraph, and title. ElasticSearch[23] operates on this in a document-oriented basis where we establish inverted indexes for both sentence and title fields. The sentence field is used for sentence-level full-text retrieval denoted as *Sentence Retrieval (SR)*. The title field signifies the primary entity described on the wiki page and serves as a field for entity-level retrieval denoted as *Entity Retrieval (ER)*. The paragraph field contains detailed info on the subject matter.

[21] BanNERD Bangla NER Annotation Guideline
[22] https://dumps.wikimedia.org/
[23] https://www.elastic.co/

In *SR*, ElasticSearch uses the Okapi BM25 retrieval algorithm (Robertson and Zaragoza, 2009) to find sentences syntactically similar to the input sentence. For *ER*, the sentences that contain the entities in the input query get a higher similarity score. The final similarity score is calculated as $S = SR + \alpha ER$ where we set $\alpha = 2$ following (Wang et al., 2022a). Based on the similarity scores, we can retrieve top-k results that are used as external contexts in our system. For the ER module, we use the gold entity labels for training and validation data. On test data, we do not use the gold labels. We first pass sentences without any entities to the *SR* module and retrieve contexts using only sentence similarity. The retrieved contexts are passed to the model and the predicted entity mentions are used in the second step for the *ER* module. We repeat this prediction-retrieval-prediction step two times following Wang et al. (2022a).

We present the two context retrieval strategies namely sentence retrieval and entity retrieval strategies in **Table 9**:

I Baselines

In this paper, we compare the *BanNERCEM* model with the following baseline models:

- **BiLSTM-CRF**: A multi-layered bidirectional LSTM network followed by a CRF decoder network.
- **BERT-CRF:** Comprised of a BERT-like encoder followed by a CRF decoder network.
- **BERT-BiLSTM**: Comprised of a BERT-like encoder followed by a multi-layered bidirectional LSTM network.
- **Banner:** Banner method (Ashrafi et al., 2020) improves on the BERT-CRF approach by using a weighted Cross Entropy loss based on the abundance/rarity of NER tags and a two-stage training mechanism to enhance performance.
- **Noisy Label:** Noisy Label (Zhou and Chen, 2021) approach tackles the noise in human annotations with an additional Kullback–Leibler (KL) divergence loss between two BERT-biLSTM-CRF models.
- **RaNER:** RaNER (Wang et al., 2022a) is a context-based NER approach that retrieves the top 10 contexts from Wikipedia Dump and concatenates it to the input sentence. We follow their official implementation[24].
- **DiffusionNER:** A novel approach that frames Named Entity Recognition (NER) as a boundary-denoising diffusion process. The method generates named entities by progressively refining noisy spans. During training, noise is added to the correct entity boundaries through a forward diffusion process, and the model learns to reverse this process to recover the boundaries. During inference, noisy spans are sampled from a Gaussian distribution and denoised using the learned reverse process to generate entities. This boundary-denoising approach enables flexible and efficient entity generation. Experiments on both flat and nested NER datasets show that DiffusionNER achieves performance comparable to or better than current state-of-the-art models. We follow their official implementation [25]
- **Binder:** Introduces a bi-encoder framework for Named Entity Recognition (NER) that uses contrastive learning to align text spans and entity types within a shared vector space. Unlike traditional sequence labeling or span classification approaches, this method reframes NER as a representation learning task, maximizing similarity between entity mentions and their types. The method demonstrates state-of-the-art performance across both supervised and distantly supervised settings on general-domain datasets (ACE2004, ACE2005, CoNLL2003) and specialized biomedical datasets (GENIA, NCBI, BC5CDR, JNLPBA). We follow their official implementation [26].
- **OpenAI GPT models:** We used GPT-4 Omni (GPT-4o) which is the current most capable model from OpenAI (OpenAI, 2024a) and gpt-3.5-turbo-0125 which is the most capable gpt-3.5 text model (OpenAI, 2024b). We have followed the prompting strategy of (Lai et al., 2023) to construct the NER recognition prompt for each dataset. We use English

[24] https://github.com/Alibaba-NLP/KB-NER
[25] https://github.com/tricktreat/DiffusionNER
[26] github.com/microsoft/binder

prompt as it is shown to perform better in multilingual setup by Lai et al. (2023). As shown in **Figure 4**, the prompt consists of three parts: *Task Description, Note* and *Demonstrations*. *Task description* explains the NER task and the NER tags. As we have experimented on various Bangla NER datasets, the NER tags vary according to each dataset. The *Note* explains the BIO format and the *Demonstrations* part contains **ten** examples containing the Input and the corresponding output to guide ChatGPT's response. The ten examples are created from each dataset's training set. Examples are chosen in such a way that every NER tag has at least one example in these ten instances. After the *Demonstrations* section, we provide a single Input for which ChatGPT has to produce the corresponding output. Due to high costs for API usage for GPT-4o, GPT-3.5 Turbo, we conduct evaluation a random 1000 split on the test set of all datasets.

```
Task Description: You are working as a named entity recognition expert and your task
is to label a given text with named entity labels. Your task is to identify and label
any named entities present in the text. The named entity labels that you will be
using are NUM (number), EVENT (event), MISC (miscellanous ner tags),
PER (person),T&T (terms and titles of different entities.), LOC (location), GPE
(geographical and political entities), UNIT (measurements of money, number, rate,
age, etc.), D&T (Data & Time), ORG (Organization).

Note: Please use BIO annotation schema to complete this task. Please make sure to
label each word of the entity with the appropriate prefix ("B" for
the first word of the entity, "I" for any non-initial word of the entity). For words
which are not part of any named entity, you should return "O". Do not provide any
explanations for the output.

Demonstrations: Optional.
[Input:['একেকটি', 'বই', '৭০', 'থেকে', '৯০', 'টাকা', 'বা', 'এর', 'চেয়েও', 'বেশি',
'দামের', 'ই'] ,
Output:['O', 'O', 'B-NUM', 'O', 'B-NUM', 'B-UNIT', 'O', 'O','O', 'O', 'O', 'O']],
[Input:['কলেজে', 'ভর্তিতে', 'কোন', 'আসন','সংকট', 'হবে', 'না', 'বলে', 'জানিয়েছে',
'শিক্ষাবোর্ড', 'ই'] ,
Output:['B-ORG', 'O', 'O', 'O', 'O', 'O', 'O','O','O', 'B-ORG', 'O']],
[Input:['আমরা', 'আমাদের', 'জাতীয়', 'স্বার্থ','রক্ষার', 'জন্য', 'লড়াই', 'করছি', ',',
'আমাদের', 'নাগরিকদের', 'ও', 'জনগণের', 'স্বার্থের', 'সুরক্ষায়', 'লড়ছি', 'ই'] ,
Output:['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O','O', 'B-PER', 'O', 'B-PER', 'O',
'O', 'O', 'O']],
[Input:['মেঘলা', ',', 'নীলাচল', ',', 'প্রান্তিক','লেক', 'শৈলপ্রপাত', ',', 'চিম্বুক',
'পাহাড়', ',',',নীলগিরিসহ', 'দর্শনীয়স্থানে', 'ভিড়', 'করছেন', 'তারা', 'ই'] ,
Output:['B-LOC', 'O', 'B-LOC', 'O', 'B-LOC','I-LOC', 'I-LOC', 'O', 'B-LOC', 'I-LOC',
'O','B-LOC', 'O', 'O', 'O', 'O', 'O']],
[Input:['এদিকে', ',', 'নতুন', 'বছর', 'উপলক্ষ্যে','দেওয়া', 'ভাষণে', ',', 'ইউক্রেন',
'সংকটের', 'জন্য','আবারো', 'পশ্চিমাদের','দায়ী', 'করেন', 'রুশ', 'প্রেসিডেন্ট', 'ই'] ,
Output:['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O','B-GPE', 'O', 'O', 'O', 'B-PER', 'O',
'O', 'B-GPE', 'B-PER', 'O']],
[Input:['আজও', 'কুয়াশার', 'চাদরে', 'ঢাকা', 'রাজধানী', 'ই'] ,
Output:['B-D&T', 'O', 'O', 'O', 'O', 'O']],
[Input:['আটকে', 'যায়', 'রাজশাহীর', 'ফজলি','আমের', 'জিআই', 'সনদ', 'ই'] ,
Output:['O', 'O', 'B-GPE', 'B-MISC', 'I-MISC', 'B-MISC', 'I-MISC', 'O']],
[Input:['ইংরেজী', 'নতুন', 'বছর', 'বরণের', 'অন্যতম','আকর্ষণ', 'যুক্তরাষ্ট্রের',
'নিউইয়র্কের','টাইসস্কয়ারের', 'নিউ', 'ইয়ার', 'ইভের', 'আয়োজন', 'ই'] ,
 Output:['O', 'O', 'O', 'O', 'O','O', 'B-GPE', 'B-GPE', 'B-LOC', 'B-EVENT', 'I-
EVENT','I-EVENT', 'O', 'O']],
[Input:['গিনিজ', 'ওয়ার্ল্ড', 'রেকর্ডসে', 'নাম', 'লিখিয়েছে', 'সংযুক্ত','আরব', 'আমিরাতের',
'বর্ষবরণের', 'আয়োজন', 'ই'] ,
Output:['B-T&T', 'I-T&T', 'I-T&T', 'O', 'O', 'B-GPE','I-GPE', 'I-GPE', 'O', 'O',
'O']],
[Input:['তবু', 'বেশ', 'লাগছে', 'বসে', 'থাকতে', 'ই'] ,
Output:['O', 'O', 'O', 'O', 'O', 'O']],

Input: ['স্বাধীন', 'বাংলাদেশে', 'মাথাপিছু', 'আয়', 'ছিলো', '৮৮', 'ডলার', 'ই']
Output:
```

Figure 4: Input prompt format for ChatGPT experiment. **Task description** describes the NER tags, **Note** describes the Output format and the **Demonstrations** show ten examples which cover all NER tags present for any specific NER dataset. The Task description and the Demontrations part vary from dataset to dataset.

J Experimental Settings

As mentioned earlier, for each respective baseline system, we use the hyperparameters specified in their respective works. As different datasets have achieved the best performance using different PLMs, we use the respective PLM used for each dataset for a fair comparison with our approach. We use the BanglaBERT (Bhattacharjee et al., 2022) model for BanNERD, Karim et al. (2019) and WikiANN (Pan et al., 2017) dataset. For MultiCoNER datasets (Malmasi et al., 2022b; Fetahu et al., 2023a), we use the XLMRoBERTa (XLM-R) (Conneau et al., 2020) model as used in RaNER model (Wang et al., 2022a; Tan et al., 2023). (Haque et al., 2023) achieved highest performance using NR/IndicbnBERT [27]. But as the resulting transformer model is now unavailable on Huggingface [28], we have used their second best-performing transformer model, SagorBERT (Sarker, 2020) for our experiments. For Naamapadam (Mhaske et al., 2023) dataset, we use bert-base-multilingual-cased [29] model.

For the BiLSTM-CRF network, we use a pre-trained Bangla Fasttext embedding trained with 20M words from the bnlp toolkit [30] as (Haque et al., 2023) achieved higher performance with the Fasttext embedding compared to other Bangla pre-trained embeddings. For the BERT-BiLSTM network, we train the whole network meaning the BERT network is trained along with the BiLSTM network following (Ashrafi et al., 2020). For all experiments, we use the AdamW optimizer (Loshchilov and Hutter, 2019) with a linear warmup-decay learning scheduler, train the models for 40 epochs without using early stopping, and set the batch size to 64. We choose the best model based on validation scores; except for MultiCoNER datasets. For other hyperparameters, we do not do any hyperparameter tuning and choose the ones used in each respective method. Following RaNER (Wang et al., 2022a), we combine the train and validation sets to fully utilize all data. On these datasets, we choose the final model for inference on the test set. As mentioned earlier, we take the top 10 retrieved contexts from the *CRM* module for our *BanNERCEM* approach. We use a

[27] https://huggingface.co/neuralspace-reverie/indic-transformers-bn-bert

[28] https://huggingface.co

[29] https://huggingface.co/google-bert/bert-base-multilingual-cased

[30] https://github.com/sagorbrur/bnlp/tree/main

single NVIDIA A40 48GB GPU for all training.

K Additional Results

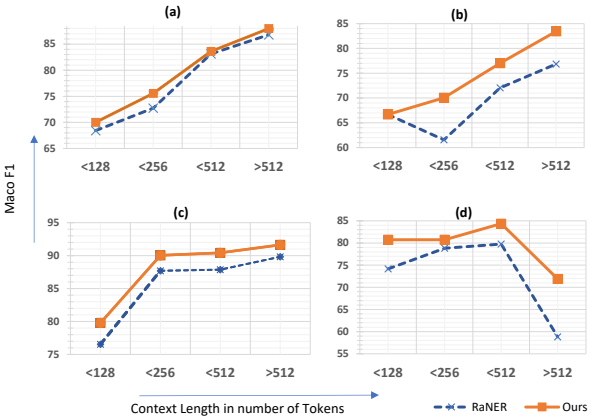

Figure 5: Comparison between **RaNER** and **BanNERCEM** (ours) on varying context lengths in (a): MultiCoNER I Malmasi et al. (2022b), (b): MultiCoNER II (Fetahu et al., 2023a),(c): BanNERD and (d): (Haque et al., 2023) datasets in terms of entity-wise macro F1 score.

LLM Name	Macro F1	Parameters (in Millions)
IndicBERT (Doddapaneni et al., 2023)	80.27	130
Vac-BERT (Bhattacharyya et al., 2023)	80.4	17
sahajBERT[31]	86.24	18
BanglaBERT (Bhattacharjee et al., 2022)	87.36	110
MuRIL (Khanuja et al., 2021)	**87.84**	256

Table 10: Comparison between various LLMs using the Banner (Ashrafi et al., 2020) approach on the BanNERD dataset. MuRIL is the largest and performs the best.

We perform an extensive analysis of our proposed approach to analyze its effectiveness in various settings. Our *BanNERCEM* approach passes each context separately to the NER model to circumvent the token length limitation of PLMs. To analyze if this is effective, we divide the NER instances according to their context lengths and report the performance of **RaNER** and our **BanNERCEM** approach in **Figure 5** on four Banla NER datasets including our BanNERD dataset. **Figure 5** shows that BanNERCEM consistently outperforms the RaNER on all context length ranges. More importantly, BanNERCEM achieves 5.72% greater average macro F1 score on these datasets than RaNER when context length exceeds 512 tokens. On average, 7.8% contexts exceed the length of 512 tokens. When ten contexts are concatenated, many contexts are being stripped to 512 tokens which underutilizes these contexts and leads to poorer performance. Performance generally increases as the context length increases showing that the model is utilizing the extra knowledge of the contexts.

Apart from the LLMs used in **Table 3**, we have also experimented with other available pre-trained Bangla LLMs using the Banner (Ashrafi et al., 2020) approach and provide their results in **Table 10**. Based on the results, sahajBERT despite its small size of 18 million parameters performs better than other bigger models like IndicBERT (130 million parameters). BanglaBERT (110 million parameters) which we use to report our results for the BanNERD dataset is in second position but less than half the size of MuRIL (256 million parameters). So, considering size and performance - BanglaBERT gives the best of both worlds.

Dataset	BanglaBERT	Reported Result
BanNERD *	89.39	89.39
Haque et al. (2023)	72.37	83.62
Karim et al. (2019)*	72.03	72.03
Malmasi et al. (2022b)	70.36	83.18
Fetahu et al. (2023b)	52.27	76.55

Table 11: Entity-wise macro f1 score of BanglaBERT model (Bhattacharjee et al., 2022) on all datasets using our proposed *BanNERCEM* approach. The $*$ sign suggests that the BanglaBERT model was used for reporting the original results. So, it is equal to the Reported result column.

We provide the entity-wise macro-f1 scores on all datasets using BanglaBERT model (Bhattacharjee et al., 2022) in **Table 11**. On MultiCoNER datasets, BanglaBERT performs worse than XLM-RoBERTa (Conneau et al., 2019) model. This might be due to XLM-RoBERTa's pretraining data as it used a huge amount of Translated Bangla data. MultiCoNER datasets' Bangla splits are translated from English which may have proved beneficial to the XLM-RoBERTa model.

It is important to note that on the MultiCoNER 2023 dataset (Fetahu et al., 2023b), the reported result in **Table 3** is for the RaNER model. U-RaNER (Tan et al., 2023) utilizes a more intricate retrieval method than RaNER incorporating Wikidata [32] alongside Wikipedia Dump and 81.60% macro f1 score on the MultiCoNER 2023 dataset which is higher than the RaNER result. The results for our *BanNERCEM* approach in **Table 3** only

[32] https://www.wikidata.org/wiki/Wikidata:Main_Page

use Wikipedia Dump for context retrieval as mentioned in **Section H**. However, as previously mentioned in **Section 6**, the external contexts passed to the underlying model in U-RaNER have not been made public. Following the methodology outlined in U-RaNER, we have tried to reproduce their retrieval system but only managed to achieve 77.65% macro f1 score using our *BanNERCEM* model on the MultiCoNER 2023 dataset which is only slightly better than before. We also trained the U-RaNER model using contexts from the RaNER retrieval system but only achieved 39.81% macro f1 score. As the retrieval system for U-RaNER was not released and we did not properly reproduce their retrieval system, we decided to not provide the results for U-RaNER mode on all Bangla NER datasets in **Table 3** and only provided the results for RaNER model.

L Data Licensing & Legality

The project, funded by a national research funding authority limits dataset usage to non-commercial and academic research. Data collection follows the fair usage policies of Facebook[33] using standard API calls and random sampling from public pages and comments. Anonymization and de-identification steps are taken to protect privacy. Personal mentions are anonymized, ensuring anonymity and de-identification following the funding body's rules.

[33] https://www.facebook.com/help/1020633957973118