IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 June 2022, accepted 19 July 2022, date of publication 1 August 2022, date of current version 4 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195236

RESEARCH ARTICLE

Compilation, Analysis and Application of a Comprehensive Bangla Corpus KUMono

AYSHA AKTHER, MD. SHYMON ISLAM, HAFSA SULTANA, A. K. Z. RASEL RAHMAN, SUJANA SAHA, KAZI MASUDUL ALAM, AND RAMESWAR DEBNATH

Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh

Corresponding author: Aysha Akther (aysha@cseku.ac.bd)

This work was supported in part by the Information and Communication Technology Division (ICT Division), Government of the People's Republic of Bangladesh (S-Number 56.00.0000.052.33.003.22-49, 12-06-2022), and in part by the Khulna University Administration (by paying the publication fee partially).

ABSTRACT Research in Natural Language Processing (NLP) and computational linguistics highly depends on a good quality representative corpus of any specific language. Bangla is one of the most spoken languages in the world but Bangla NLP research is in its early stage of development due to the lack of quality public corpus. This article describes the detailed compilation methodology of a comprehensive monolingual Bangla corpus, KUMono (**K**hulna **U**niversity **Mono**lingual corpus). The newly developed corpus consists of more than 350 million word tokens and more than one million unique tokens from 18 major text categories of online Bangla websites. We have conducted several word-level and character-level linguistic phenomenon analyses based on empirical studies of the developed corpus. The corpus follows Zipf's curve and hapax legomena rule. The quality of the corpus is also assessed by analyzing and comparing the inherent sparseness of the corpus with existing Bangla corpora, by analyzing the distribution of function words of the corpus and vocabulary growth rate. We have developed a Bangla article categorization application based on the KUMono corpus and received compelling results by comparing to the state-of-the-art models.

INDEX TERMS NLP, Bangla corpus, N-gram, Zipf's law, article categorization.

I. INTRODUCTION

In this era of real-time multimedia data generated in digital and analog sources all around us, natural language processing (NLP) is becoming more and more critical to make sense of data. To retrieve the true essence of knowledge from this enormous data, we need essential resources such as a corpus to understand the words and their relationships. Applications of corpora have been extended to all areas of computational linguistics and natural language processing (NLP) especially for generating training sets for language modeling and machine learning.

English and some other European languages have long developed their own standards and different types of special corpora and various language processing tools. Bangla is the seventh most spoken language and is used every day by more than 250 million people around the world, such as the primary language in Bangladesh and the second language in India [1]. Yet research on Bangla language processing is way behind the other languages with rich language resources. Very few notable works on language resource building in Bangla have been conducted. In this research, we aim to develop a monolingual Bangla corpus based on an automatic collection of very large-scale data from a wide variety of domains and styles.

In linguistics, the term corpus generally refers to a collection of texts upon which some general linguistic analysis can be conducted. More specifically according to modern linguistics corpus should represent several properties such as machine-readable form, sampling and representativeness, finite size, and means a standard reference for the language variety [2]. Corpus is used as an "example bank" by many linguists as empirical support for their hypothesis. It also serves as a means for a quantitative study like frequency information of words and phrases. A corpus may provide metadata such as genre, temporal and spatial information

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna D'Ulizia.

about the origin of the texts, and based on a quantitative study it can provide similarities and dissimilarities between different types of texts. Quantitative information extracted from the corpus is applied by computational linguistics and by theoretical linguistics as well.

Corpus can be primarily divided into two genres: written corpus and spoken corpus. Texts of a written corpus comprised of language data collected from various written, printed, published, and electronic sources. Brown Corpus is the first major structured corpus of varied genres of American English [3]. Texts of spoken corpus comprised of speech data in written form by transcription. London-Lund Corpus is an example of spoken corpus [4]. A sample or reference corpus of a language should be a reliable repository of all the features of the language. A sample corpus should consist of several hundreds of millions of words from the language of spoken and written, public and private, informative and fictional, etc. of a society. A corpus can be designed by including the time dimension as a design feature [5]. A corpus can also be bilingual or multilingual. Canadian Hansard is the first example of parallel or bilingual corpus [2]. A corpus is specialized when it is designed based on specific research goals. Cambridge and Nottingham Corpus of Discourse in English (CANCODE) [6] and Michigan Corpus of Academic Spoken English (MICASE) [7] are examples of the specialized corpus. Some corpus comprised of important collections and extraordinary works of a language rather than gathering ordinary language usage phenomena such as the works of Shakespeare is another example of special corpus [2].

Corpus-based research has been carried out by linguistic researchers from the previous century. Biblical concordances represent the first significant examples of corpus-based research with linguistic associations [8]. Corpus is also applied for the compilation of grammar for English and some other European languages. [9] and [10] are examples of large-scale corpus-based reference grammars. Corpora are now commonly used as the basis for creating dictionaries [3]. The empirical analysis of language provided by Corpus Linguistics (CL) is being used in broader research areas such as language teaching and learning, discourse analysis, literary stylistics, forensic linguistics, pragmatics, speech technology, sociolinguistics, and health communication, among others. Major publishers such as Oxford University Press, Cambridge University Press, Pearson-Longman, Collins-COBUILD, and Macmillan all closely guard multi-million-word corpora as language teaching material and regularly launch new materials which are corpus informed [3].

In this era of enormous data all around us, corpora have become a compulsory component for applications of Natural Language Processing (NLP) and computational linguistics. Probabilistic model-based approaches are mainly used for building language models and machine learning algorithms for language processing. Probabilistic models yield better results even for large-scale noisy data rather than carefully selected limited balanced data [11]. Moreover, the Bangla language has distinct linguistic features from other languages. Bangla has 11 vowels and 39 consonants. The use of vowel allographs, clustered letters in words, word formation rules, and sentence structure contribute to making Bangla distinct. Statistical analysis of a very large-scale Bangla corpus would produce reliable results for distinct linguistic features of the Bangla language. So in this research, we emphasize building a large-scale corpus of the Bangla language. We have developed a monolingual written corpus of the Bangla language consisting of several hundred million words. We found various Bangla websites as the source of collecting a large amount of text or data.

Our corpus contains more than 353 million word tokens and 1,686,270 unique word tokens from 1,292,527 online articles. We named the developed corpus as KUMono corpus. We hope language researchers would find this corpus useful in the study of both theoretical and computational linguistics.

The rest of the article is organized as Section II describes the history of building corpus, traditional corpus design practices, state-of-the-art corpora building approaches in various languages, and observations about existing Bangla corpus building approaches. Section III states the scope and objective of this research, Section IV presents the step-by-step process of compilation, structuring, and documentation process of the Bangla corpus, KUMono. Next, several word-level and character-level linguistic phenomena are analyzed on the KUMono data in Section V followed by a number of evaluations of the quality of the developed corpus in Section VI. Section VII explains article categorization as an application of the KUMono corpus. Finally, this article concludes with some other possible applications and future works on KUMono in Section VIII.

II. BACKGROUND AND RELATED WORKS

This section briefly discusses history of building corpus, mentions some influential English corpora, some non-English corpora, and some important corpus building approaches that are previously done in Bangla language. Finally, we pointed out our observations of previous Bangla corpora.

A. HISTORY OF BUILDING CORPUS

Historically corpora can be divided into two types: pre-electronic corpora and electronic corpora. In the pre-electronic era, corpus creation and analysis process was time-consuming and required tedious manual labour. The most significant pre-electronic corpus was the Survey of English Usage (SEU) Corpus of spoken and written texts whose compilation began in 1959 at University College London [3]. Although the tedious manual analyses associated with pre-electronic corpora are now seemed needless, these corpora had played an important role in the development of corpus linguistics as a field of research. Works from 1950-1970 on natural language texts were mainly dependent on hand-crafted rule-based systems. But rule-based systems are not entirely suitable for natural language processing since they are not easily extensible to large-scale texts. Texts from

natural language contain many unknown words and are very ambiguous which is impossible to handle with rule-based systems. The need for large-scale machine-readable corpus becomes obvious for research on natural language processing. A real breakthrough in compilation and analysis of corpora began with the access to machine readable texts which could be stored, transported, and analyzed electronically [2].

B. ENGLISH LANGUAGE CORPUS

In the 1970s and 1980s, there was an explosion in the quantity and diversity of texts compiled and analyzed by computer. Brown Corpus, the LOB Corpus, and the London-Lund Corpus are considered as examples of first-generation computerized English corpora [8]. Though they were not the only early computer corpora, they were the most influential ones. As soon as the world wide web began to exploit as the primary source of data, billion words corpus become possible. The Cambridge English Corpus (formerly known as Cambridge International Corpus) [12] and The Oxford English Corpus (OEC) [13] both are multi-billion word text corpora of 21st-century English. In this modern era of speed and efficiency, young scholars are less willing to spend years just collecting materials from manuscripts [2]. Now more emphasis is given to coverage of more extensive text and enhancing the depth and breadth of analysis. Now the application area of new corpora with more accurate and multifaceted annotation have been extended to discourse and pragmatic analysis [2]. The performance of Natural Language Processing applications based on probabilistic models highly depends on the size of the corpora [11]. Some recent corpus based research works are described in the following paragraphs.

C. NON-ENGLISH LANGUAGE CORPUS

There are numerous examples of non-English corpora in many languages such as Spanish, Swedish, Portuguese, Bulgarian, Mandarin, Japanese, etc. We only mention a few non-English corpora here. In [14] authors describe the work carried out on the EMILLE Project (Enabling Minority Language Engineering), which was undertaken by the Universities of Lancaster and Sheffield. The EMILLE corpus is a collection of fourteen South Asian languages' monolingual corpora with more than 96 million words. It also includes a parallel corpus of English and five of these languages. The paper explains the collection, storing, and processing steps of the corpus in detail. The corpus also includes a parts-of-speech tagged annotated component for one of the languages. Uppsala written corpus of student writings is another example of a recent corpus of Swedish texts [5]. The Uppsala Corpus of Student Writings consists of several years of Swedish texts produced as part of a national test for students aged from 9 - 19 years. The corpus consists of 2,500 texts and 1.5 million tokens. Parts of the texts have been annotated on several linguistic levels. Research work [11] describes the compilation and annotation of Bulgarian National Corpus (BulNC) composed of 979.6 million words. The main emphasis of this research was to increase the size of the corpus rather than following the traditional corpus designing approach. The corpus contains a Bulgarian part and a Bulgarian-English parallel corpus part. The Bulgarian part contains annotations of several levels such as morphosyntactic tagging, lemmatization, word-sense annotation, annotation of noun phrases, and named entities.

D. BANGLA LANGUAGE CORPUS

In recent years, a few Bangla language researchers have demonstrated a keen interest in Bangla corpus development and Bangla text analysis. Some notable progress has been documented in corpus creation in [15]–[18], [19], [20] and knowledge engineering on Bangla language in [21]–[24], etc. The first electronic Bangla corpus was constructed by the Central Institute of Indian Languages (CIIL) from 1991 to 1995 [25]. CIIL Bangla corpus had three million words. In [25], the authors analyzed the linguistic features of the Bangla language based on the CIIL Bangla corpus of three million words. But Indian Bangla language and the Bangladeshi Bangla language has some differences in terms of phonetic structure and writing style. So every analysis that is true for the Indian Bangla language might not be true for the Bangladeshi Bangla language. Moreover, a corpus consisting of 3 million words is not enough to represent every linguistic feature of any language correctly.

Authors of [16] address the issue of automatic Bangla corpus creation by focusing on language detection and Unicode conversion of retrieved text. In [17] authors aimed to develop a standard corpus on the Bangla language and collected data from different web domains that contain 7.6 million words (tokens) in total and 285,496 unique word types. They have also performed a few analyses of language phenomena on the corpus. In [15] authors presented the compilation methodology of a news corpus of a specific newspaper of a certain year that contains 18.067 million word tokens and 386,639 unique word types. They also have conducted some statistical analyses based on the word frequency count of the corpus. In [18] authors presented construction methodology and some statistical analysis of a Bangla corpus composed of 27 million words from six domains. The authors claimed the corpus as a reference corpus for the Bangla language. They have also presented several results of the study of features of the Bangla language based on empirical analysis of the corpus. In [20] a new corpus NHMono01 consisting of 100,142,522 tokens was developed. Their developed spell and grammar checker application of the corpus demonstrated some excellent results.

E. OBSERVATIONS ABOUT EXISTING BANGLA CORPORA

We have several observations of previous Bangla corpus building approaches. We tried to incorporate the knowledge of the observations in our proposed Bangla corpus building approach. Our observations from previous Bangla corpus building approaches and how our developed corpus incorporates the knowledge are pointed out below.

- *Corpus Building Strategy:* Most of the previous Bangla corpora compilation was done manually or semi-automatically. Manual approaches are not suitable for the compilation of very large scale data. With the increased development of language technologies, applications of corpora as a resource for training dataset for machine learning and language modeling is inevitable. It has been proved that a large dataset performs more reliably in probabilistic machine learning than a smaller dataset even if it is noisy [11]. We attempted a fully automatic data collection approach for developing a monolingual Bangla corpus. Automatic collection of texts facilitates dynamic enlargement, electronic storage and efficient management of corpus data.
- *Small Corpus Size:* To the best of our knowledge, our developed corpus is the largest monolingual corpus in Bangla. It is many times larger than other available corpora in Bangla language both in terms of total word count and of unique word count (presented in Table 2). Probabilistic machine learning algorithms' performance highly depends on dataset size. The performance of machine learning algorithms for article categorization presented in section VII proves this statement too.
- *Small Lexicon Size:* According to corpus linguistics, there is no specific definition for optimal corpus size. Generally, optimality of corpus is ensured by adequate coverage of lexical diversity that is estimated by word-stock of the corpus [11]. The lexicon of our developed corpus is richer than all other available corpora of Bangla language. Although limited sized corpus is still appropriate for domain specific applications of corpus, the large collection of lexicons from diverse domains facilitates the applications of corpus in many fields where lexical statistics and diversity are required.
- *Unavailability of N-Grams:* Unlike other available corpora of Bangla language we have extracted word N-grams (N = 1,2,3) from our developed corpus. Various N-gram based applications such as spelling error detection and correction, grammatical error detection, information retrieval, query expansion, dictionary look-up, text compression, language identification etc. could be facilitated by extracted N-grams.
- *Meta Tagging of Corpus:* We have implemented article level tagging in our proposed corpus. Several types of information about each article are extracted and stored as article level tags. This meta-tagging approach facilitates the implementation of various corpus-based applications. In this article, we have implemented article categorization as an application of our developed corpus. All the previous Bangla article categorization approaches except [26] are based on a small dataset having a limited number of samples per category. The performance of machine learning algorithms is highly dependent on the number of samples. Our developed machine learning based Bangla article or document categorization experiment showed outstanding performance.

III. SCOPE AND OBJECTIVE OF PROPOSED WORK

Bangla is the seventh most spoken language of the world as well as rich in vocabulary and morphological variation, yet labeled as low resource language in linguistic research. This paper attempts to fill this gap by building a large scale quality Bangla corpus. The following section describes the scope and objective of this research work.

- Firstly, we contribute to resource building by developing a very large Bangla corpus with known statistics for natural language processing. In this paper, we have compiled a Bangla corpus KUMono by collecting Bangla text data from 18 major categories of text from approximately 1.3 million articles. The KUMono corpus contains more than 353 million word tokens and 1,686,270 unique word tokens. We hope this large scale corpus will perform better as a training dataset in probabilistic models.
- Secondly, we envision to make the corpus available online and freely accessible for research. We strongly believe that public access to this large-scale KUMono corpus will encourage Bangla text research in the intellectual community. The text collection process of the corpus is automatic to ensure dynamic enlargement and efficient management.
- Thirdly, we compiled important Bangla N-grams from the corpus and conducted different word and character level language profiling. We believe that the extracted statistical linguistic features of the corpus will facilitate in the development of effective Bangla language related applications.
- Fourthly, before using the developed corpus as a training dataset for different types of linguistic applications, it is obvious to assure the quality of the corpus. So we assessed the quality of the corpus using different corpus assessment techniques and compared the assessment results with existing Bangla corpora.
- Finally, to assess the performance of the corpora as a training dataset, we have implemented article categorization as an application of the corpus and compared the results with state of the art categorization results.

IV. DEVELOPMENT OF MONOLINGUAL BANGLA CORPUS KUMono

With the availability of web data and rapid development of technologies, traditional carefully selected fixed sized corpora became less useful. Traditional corpora are good for finding collocations and concordances but today's major applications of corpora include training of machine learning algorithms and developing language models. Probabilistic models based on a larger amount of data yield more reliable models even if they are noisy than smaller-sized datasets [11]. The optimal corpus size that will obtain the core vocabulary and ensure lexical diversity of any language is not determined in any language. Now researchers are mainly focusing on collecting an extensive amount of text and performing

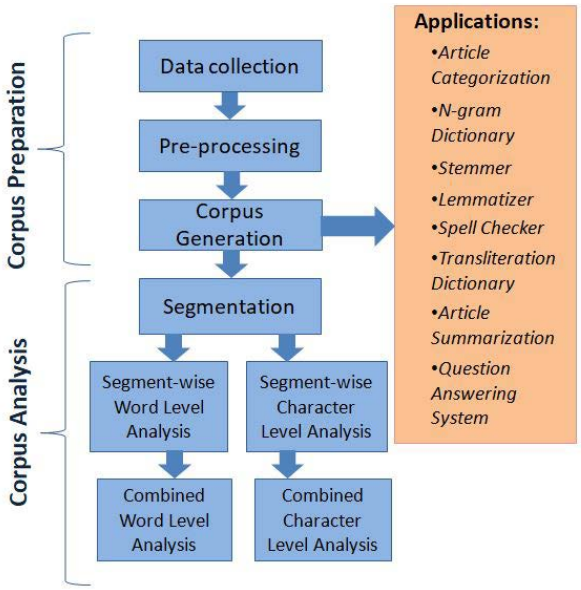

FIGURE 1. Corpus preparation and analysis diagram.

a detailed and comprehensive analysis of the text. In this study, we focus on developing a very large Bangla corpus. To develop a rich public Bangla corpus, KUMono,[1] we have crawled through online Bangla websites for data collection. We have collected articles with UTF-8 encoding. The total procedure of corpus generation and analysis is depicted in Figure 1 diagrammatically.

A. DATA COLLECTION

We have collected Bangla text from various online Bangla websites. We found the newspaper sites as the most useful sites as they are updated daily with new stories and contain managed article archives. Generally, a handful of intellectual people are engaged in the process of writing, editing, reviewing, categorizing, and publishing every online newspaper article. We have developed a specific crawler for each website to extract data. Each crawler has three segments: find main page URLs, find all unit page URLs from each main page URL and finally crawl electronic texts from specific unit pages. The pseudo-code for the designed crawler for collecting texts is shown in Algorithm 1. We store the collected texts from each article in a CSV file format as *Title*, *Author*, *Content*, *Date*, *Category*. Finally, we have developed our corpus from these articles after several steps of pre-processing as described in the next sub-section. The total number of articles in the developed corpus at the time of this reporting is 1,292,527 (approximately 1.3 million). We have crawled data simultaneously from several domains for about 4 months at a stretch. Crawling data from the web is very time consuming. So we have used both local machines (IDE: PyCharm) and Google Colab for crawling texts from the web. About 42% (540,195 articles) of data was crawled using local machine and about 58% (752,332 articles) data was crawled using Google Colab. More than 30 persons were involved in the data collection, cleaning, categorization, and verification process which can be termed as semi-automatic. Continuous Internet connection is a basic requirement for crawling data from the web. There remains a chance of data redundancy due to the internet connection interruption at some time.

We have extracted several extralinguistic metadata about the text during the crawling period. The metadata we extracted and stored during the data collection period is: *Title*, *Author*, *Content*, *Date*, *Category*. We consider these metadata as article level tagging. Most of the articles from newspaper sites and blogs were already categorized. For further quality assurance, we manually rechecked the automatically extracted tags. More than 30 persons were involved in the rechecking process. Each article was checked by two persons and if any confusion aroused then it was checked again by a third person. The final category was selected based on a majority vote. The total data collection and article tagging process took six months. Figure 2 presents an example annotation of the corpus text metadata. In Figure 2 the nodes represent the categories of metadata and lines or arcs represent the relationship between nodes.

Algorithm 1 KUMono_Crawler

```
base address ← domain link
Generate category-wise main page links using base address.
Find the main page URLs by appending date with the category-wise main page links
Store the main page URLs into a list named main_links.
Open a CSV file named main_pl in write mode.
for link in main_links do
    main_pl.writerow(link)
end for
ReadCSV ← CSV.reader(main_pl)
for main_url in ReadCSV do
    page ← requests.get(main_url)
    soup ← bs4(page.content, 'html.parser')
    Divs ← soup.find_all(div, class)
    Open Unit_links.csv in write mode
    for article in Divs do
        a_tag ← article.find(a)
        news_link ← a_tag(href)
        Unit_links.writerow(article_link)
    end for
end for
Read the Unit_links file as unit_page_urls
for url in unit_page_urls do
    title ← article_title(url)
    author ← article_author(url)
    content ← article_content(url)
    date ← article_date(url)
    category ← article_category(url)
    output ← [title, author, content, date, category]
    append output to CSV file final_result
end for
Output: In final_result CSV file, each row is the required result having title, author, content, date & category.
```

The final compiled corpus contains 353,547,583 (353.55 million) tokens from 1,292,527 (approximately 1.3 million) articles of 18 major text categories. On average, each article contains 273.53 words. The corpus has a

[1] KUMono Github Data, https://github.com/dgted/BNC

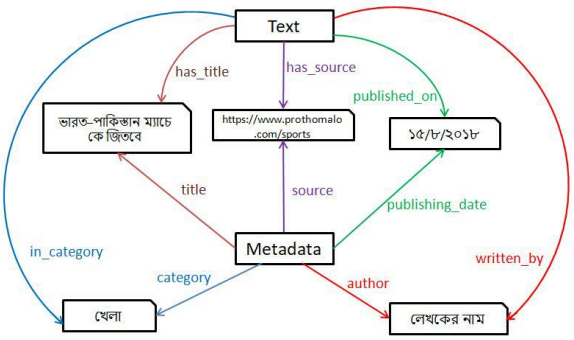

FIGURE 2. Example annotation of corpus text metadata.

wide variety of Bangla texts in different categories and each category has many sub-categories. In the corpus, we got 514 different sub-categories as different publishers categorize articles in their own way. To generalize it, we have taken similar sub-categories to map into a single one. The generalized categorization process was manually done by the involved volunteers. Finally, we have categorized all articles from the corpus into 18 categories which is shown in Table 1. When any article's category does not match with any of the major categories then we checked with the sub-categories. If the article's category matched with any of our sub-category then the article's category is tagged as the category name the sub-category belongs to. For space limitations, in Table 1 only the major sub-categories of the categories are shown. The distribution of categories in the percentage of total articles is also presented in the mentioned table. Texts from different categories and sub-categories would ensure texts genres variability and also articles from different domains of different authors ensure language usage variability. Table 1 also represents category wise count of total word, unique word, and type-to-token ratio information. The National category contains the highest number of total words but the lowest type-to-token ratio. International, Politics, and Economics categories also have a low type-to-token ratio. On the other hand, Literature, Agriculture, Environment, and Religion have high type-to-token ratio. The vocabulary of categories with high type-to-token ratio is rich with word variations. In total, our developed corpus contains 1,686,270 unique tokens from different genres of Bangla. It is difficult to maintain the balanced criteria of a corpus for a dynamically evolving corpus. Rather a balanced sub-corpus can be developed by extracting data based on domain, genre, author, time-domain etc.

A comparison table of Bangla corpora based on size is presented in Table 2. Our developed corpus is many times larger than the largest corpus among other Bangla corpora and captures more vocabulary than all other corpora.

B. DATA PRE-PROCESSING

In the pre-processing phase, we conducted data cleaning, tokenization and stop words removal. After cleaning and tokenization, we removed stop words from one version of the corpus and kept the stop words in another version.

1) DATA CLEANING AND TOKENIZATION

In the tokenization phase, the texts are split into sentences and the sentences are split into words. As in Bangla language, words are normally separated by white spaces or by punctuation marks so tokenization is comparatively easy in Bangla than in some other languages where white space does not resemble word boundary. In tokenization phase, we consider punctuation marks, brackets, numbers and hyphens as word boundary.

All the collected articles are stored in a CSV file. The size of the CSV file is around 6.00 GB. Pre-processing such a big file all at once is difficult. To perform the pre-processing and analysis efficiently we divided the whole corpus into 33 segments and performed pre-processing and analysis on each segment separately and finally, we combined the results of all segments. After cleaning with the python's built in Bangla text cleaning library, we found some punctuation and special symbols that were still present in the pre-processed corpus. For better cleaning of the dataset we have developed a python script that is also used to find the unique tokens. Our developed cleaner can recognize more symbols. For example, the builtin Python's script extracted 84542 unique tokens from a specific sub-dataset, and our improvised cleaner identified 245 unnecessary tokens. Overall, our improvised cleaner identified unnecessary 1,09,376 unique tokens from the total KUMono dataset. After punctuation removal and tokenization we got 353,547,583 word tokens and 1,686,270 unique word tokens for the entire KUMono corpus. We didn't remove numbers because some numbers have great significance such as ১৯৭১, ৬ দফা, ৭-ই মার্চ, ১৭-ই মার্চ, ১৬-ই ডিসেম্বর, etc. Our improvised cleaner played important role to develop a good quality corpus.

2) STOP WORDS REMOVAL

The words those are less significant in a document but occur frequently (e.g. অবশ্য, অথবা, অনেকে, এবং, অথচ, অন্য, আছে, ও, অন্তত, আবারো etc.) can be considered as stop words. Stop words are also called as function words of a language. Stop words or function words does not contribute to the meaning of text. So they have little value in information retrieval or knowledge engineering. So we have decided to remove these type of words from one version of our corpus to discover the content words. We have created a dictionary having 1409 such stop words or function words and removed those words from the corpus in one version. Our intention was to discover the content words and *N-Grams* from the most frequent list. After the stop word removal process, the number of unique tokens became 1,684,861 in one version of the corpus.

V. STATISTICAL ANALYSIS OF KUMono BANGLA LANGUAGE PROFILE

We have analyzed various linguistic phenomena of Bangla both at word-level and character-level based on our developed corpus. We also have extracted unigrams, bigrams,

TABLE 1. Distribution of crawled Bangla newspaper categories and sub-categories.

Category	Sub-Category	%	No. of Total Words	No. of Unique Words	Type-Token Ratio
National	Administration, Government, Capital, Country and so on	39.40	131,149,779	1,016,110	0.77
International	Asia, Europe, Africa, America, UK, Canada and so on	10.25	44,711,496	834,772	1.86
Entertainment	Bollywood, Hollywood, Music, Cinema and so on	6.31	18,073,707	446,630	2.47
Politics	Awami-League, Bnp, Diplomacy, Parliament and so on	4.64	17,655,441	283,457	1.61
Economics	Bank, Business, Trade, Share-Market, Industry and so on	3.70	14,388,370	173,309	1.20
Education	Campus, Admission, Scholarship, Notice and so on	2.04	7,627,963	176,757	2.32
Crime	Murder, Snatching, Robbery, Theft, Rape and so on	0.82	3,639,741	97,565	2.68
Sports	Cricket, Football, IPL, CPL, Tennis, Athletics and so on	11.00	39,060,283	529,015	1.35
Lifestyle	Beauty, Fashion, Food, Relationship, Job and so on	1.54	5,657,773	161,697	2.86
Technology	Science, IT, Development, Gadget, Invention and so on	1.71	6,963,157	139,689	2.01
Accident	Road accident, Clinical negligence, Injury and so on	0.37	858,933	49,015	5.71
Environment	Weather, Disaster, Climate, Plant, Animal and so on	0.16	725,481	45,807	6.31
Literature	Art, Culture, Novel, Poet, Drama, Civilization and so on	0.21	1,181,765	90,662	7.67
Health	Medical tips, Covid-19, Virus protection and so on	1.64	4,049,185	108,252	2.67
Law	Court, Justice, Law & crime, Human rights and so on	1.59	4,255,635	93,293	2.19
Religion	Islami-world, Ramadan, Ijtema, Questions and so on	0.30	1,083,908	64,665	5.97
Agriculture	Farming, Horticulture, Cultivation, Guideline and so on	0.11	516,216	34,215	6.63
Opinion	Editorial, Interview, Social-media and so on	14.21	51,948,750	1,023,392	1.97

TABLE 2. Comparison between KUMono and other contemporary Bangla corpora.

	Developed KUMono	NHMono01 [20]	SUMono [18]	Prothom-alo [15]	BDNC01 [27]	CIIL [25] [19]
Corpus Size (No. of words)	**353,547,583**	100,142,522	27,118,025	18,100,378	11,000,000	3,044,573
Vocabulary size (Unique words)	**1,686,270**	1,043,106	571,572	384,048	310,483	190,841

and trigrams from the corpus before and after removing stop words. An n-gram is a contiguous sequence of n items from a given sample of text. We extracted those n-grams from the pre-processed data. In the following sub-sections, the word level and character level characteristics of the Bangla language are analyzed and discussed step by step.

A. WORD LEVEL ANALYSIS

1) TOP 20 NGRAMS OF THE CORPUS

We have extracted the most frequent n-grams of the corpus. The top 20 unigrams, bigrams, trigrams of the corpus are listed in Table 3. The most frequent words of any language are always the function words or stop words. They are small in length and belong to a closed set of words that mark grammatical structure rather than referring to something meaningful. In any reasonably sized corpus, function words are the top frequent words. From Table 3 it is also evident. They are important for assessing the quality of a corpus. We observed that the function words cover 35.57% of the total tokens of our corpus. But a corpus has some other usage such as information retrieval that needs to ignore the function words. To retrieve the content words of the corpus, we have removed the stop words from one version of the corpus. In Table 4, the top 20 frequent unigrams, bigrams, and trigrams are listed after removing the stop words.

2) UNIGRAM

In the fields of computational linguistics and probability, an n-gram consists of a single item from a given sample of text or speech is known as a unigram. In this stage, unigrams have been extracted for non-stemmed words from the pre-processed data. We have extracted unigrams for the complete dataset. For computational efficiency, the corpus is divided into several sub-datasets. Unigrams are extracted from each subset and finally results from all subsets are merged. For the KUMono dataset, we have recorded the total number of lexicons (tokens/unigrams), the total number of unique unigrams, an alphabetically sorted dictionary of unique unigrams, a frequency-based (descending order) dictionary of unique unigrams.

All unique unigrams of all sub-datasets are merged to make a final unique unigram dictionary for the developed KUMono corpus with the corresponding frequency count. The developed KUMono corpus has a total of 1,686,270 unique tokens (unigrams). From our unigram dictionary, we have shown the frequency-wise top 20 unigrams in Table 3. All these unigrams are stop words. Among them, Bangla letter ও has the highest frequency 4,074,258. Table 4 presents the top 20 unigrams after removing stop words.

3) BIGRAM

A bigram is a sequence of two adjacent words from a given sample of text. A bigram is an n-gram for n = 2.

TABLE 3. Top 20 frequent N-grams without removing stop words in KUMono corpus.

Rank	Unigram	Frequency	Bigram	Frequency	Trigram	Frequency
1	ও	4,074,258	করা হয়েছে	512,798	প্রধানমন্ত্রী শেখ হাসিনা	77,031
2	করে	2,955,083	করা হয়	466,507	এসব কথা বলেন	45,244
3	থেকে	2,382,759	তিনি বলেন	354,889	বঙ্গবন্ধু শেখ মুজিবুর	42,970
4	এ	2,339,141	করা হবে	203,274	গত ২৪ ঘণ্টায়	38,351
5	করা	2,088,253	আওয়ামী লীগের	183,085	তিনি আরও বলেন	33,050
6	না	1,959,425	করতে হবে	182,224	সূত্রে জানা গেছে	30,040
7	হয়েছে	1,944,833	এ সময়	180,487	মেডিকেল কলেজ হাসপাতালে	26,486
8	এই	1,809,507	সাধারণ সম্পাদক	180,420	এ ঘটনা ঘটে	25,639
9	হয়	1,794,023	এর আগে	1801,63	বলে জানা গেছে	23,318
10	তিনি	1,769,081	জানা গেছে	177,869	থানার ভারপ্রাপ্ত কর্মকর্তা	20,603
11	বলেন	1,700,188	শেখ হাসিনা	140,871	উদ্ধার করা হয়	20,557
12	এবং	1,643,545	করা হচ্ছে	140,433	লীগের সাধারণ সম্পাদক	18,747
13	হবে	1,495,761	উপস্থিত ছিলেন	135,435	তিনি আরো বলেন	18,110
14	তার	1,408,321	হতে পারে	132,908	উপজেলা নির্বাহী কর্মকর্তা	16,964
15	জন্য	1,295,255	প্রধানমন্ত্রী শেখ	125,980	ভারপ্রাপ্ত কর্মকর্তা ওসি	16,076
16	নিয়ে	1,212,769	এ বিষয়ে	120,690	শেখ হাসিনা বলেন	15,302
17	এক	1,173,991	এর মধ্যে	114,745	হাসপাতালে ভর্তি করা	11,765
18	করেন	1,081,056	বলা হয়েছে	107,797	এ কথা বলেন	11,685
19	করতে	1,045,130	কোটি টাকা	105,409	আওয়ামী লীগের সাধারণ	11,632
20	একটি	1,021,480	হবে না	102,040	জাতির পিতা বঙ্গবন্ধু	11,560

TABLE 4. Top 20 frequent N-grams after removing stop words in KUMono corpus.

Rank	Unigram	Frequency	Bigram	Frequency	Trigram	Frequency
1	হাজার	903,948	আওয়ামী লীগের	183,085	প্রধানমন্ত্রী শেখ হাসিনা	77,031
2	সময়	727,209	সাধারণ সম্পাদক	180,420	বঙ্গবন্ধু শেখ মুজিবুর	42,970
3	বাংলাদেশ	711,759	শেখ হাসিনা	140,871	গত ২৪ ঘণ্টায়	38,351
4	টাকা	632,597	প্রধানমন্ত্রী শেখ	125,980	মেডিকেল কলেজ হাসপাতালে	26,486
5	পুলিশ	569,110	কোটি টাকা	105,409	থানার ভারপ্রাপ্ত কর্মকর্তা	20,603
6	দেশের	505,593	লাখ টাকা	89,666	লীগের সাধারণ সম্পাদক	18,747
7	লাখ	493,674	কোভিড ১৯	83,563	উপজেলা নির্বাহী কর্মকর্তা	16,964
8	ঢাকা	451,498	হাজার টাকা	78,938	ভারপ্রাপ্ত কর্মকর্তা ওসি	16,076
9	করোনা	447,944	মেডিকেল কলেজ	59,894	আওয়ামী লীগের সাধারণ	11,632
10	বছর	445,476	২৪ ঘণ্টায়	49,322	জাতির পিতা বঙ্গবন্ধু	11,560
11	সরকার	403,194	বঙ্গবন্ধু শেখ	46,974	শেখ মুজিবুর রহমানের	9,812
12	শেখ	402,965	হাসপাতালে ভর্তি	46,063	সিভিল সার্জন ডা	7,711
13	ইসলাম	401,424	সংবাদ সম্মেলনে	43,399	করোনা ভাইরাসে আক্রান্ত	7,694
14	কোটি	366,245	আক্রান্ত হয়ে	38,142	প্রধান অতিথির বক্তব্যে	7,566
15	জাতীয়	364,099	উপজেলা স্বাস্থ্য	37,834	উপজেলা নির্বাহী অফিসার	7,131
16	বিরুদ্ধে	361,196	পুলিশ সুপার	36,591	সাকিব আল হাসান	6,403
17	বাংলাদেশের	350,554	সংসদ সদস্য	34,417	উপজেলা স্বাস্থ্য কমপ্লেক্সে	5,373
18	প্রধান	338,160	করোনায় আক্রান্ত	34,185	থানার অফিসার ইনচার্জ	4,833
19	সাধারণ	337,959	থানার ওসি	33,078	লাখ টাকার শেয়ার	3,187
20	জেলা	337,385	শেখ মুজিবুর	29,026	টাকার শেয়ার লেনদেন	2,867

The frequency distribution of every bigram in a text is commonly used for simple statistical analysis of text in many NLP and computational linguistics applications. Bigrams have been extracted from non-stemmed words. For each sub dataset, we have recorded the number of all bigrams, the total number of unique bigrams, an alphabetically sorted dictionary of unique bigrams, and a frequency-based dictionary for unique bigrams (in descending order).

All unique bigrams of all sub-datasets are merged to make a final unique bigram dictionary for the developed KUMono corpus with the corresponding frequency count. The developed corpus has a total of 89,518,234 (89.51 million) possible unique bigrams. Among them করা হয়েছে is the bigram having the highest frequency 512,798.

To accept two consecutive words as a bigram, the total number of times the consecutive words appeared together is calculated. We have considered 4 threshold frequencies: 20, 50, 100, and 200 to extract bigrams from the KUMono corpus. Threshold frequency refers to the limit or boundary of accepting or rejecting a bigram from all bigrams. Threshold

TABLE 5. Threshold frequency wise Bigram.

Threshold frequency	No. of All Bigrams	No. of Unique Bigrams
20	1,982,599	342,388
50	690,192	122,353
100	299,061	62,543
200	125,814	34,873

TABLE 6. Threshold frequency wise Trigram.

Threshold frequency	No. of All Trigrams	No. of Unique Trigrams
20	682,795	214,598
50	185,629	69,074
100	78,074	39,281
200	38,479	26,364

frequency 20 means if a bigram frequency (count) is at least 20 then it is accepted, otherwise rejected. Using threshold frequency 20, the developed corpus has a total of 1,982,599 (almost 2 million) bigrams and 342,388 unique bigrams. In Table 5, we have shown the information of bigrams using four threshold frequencies.

4) TRIGRAM

A trigram is a sequence of three adjacent elements from a given sample of text or speech. A trigram is an n-gram for $n = 3$. The frequency distribution of trigrams in a text is commonly used for simple statistical analysis of text in many applications like bigrams. Like bigrams, we have recorded the number of all trigrams, the total number of unique trigrams, an alphabetically sorted dictionary of unique trigrams, and a frequency-based dictionary for unique trigrams(in descending order) for all sub-datasets. All unique trigrams of all sub-datasets are merged to make a final unique trigram dictionary for the developed KUMono corpus with the corresponding frequency count. The developed KUMono corpus has 165,458,839 (165.46 million) unique trigrams. Trigrams have been also extracted using same threshold frequencies like bigrams. Using threshold frequency 20, the developed corpus has a total of 682,795 trigrams and 214,598 unique trigrams. In Table 6, we have shown the number of trigrams using four threshold frequencies.

5) AVERAGE WORD LENGTH

There are about 1887,590,281 characters excluding spaces and punctuations in the corpus with an average of 5.34 letters per word. In ordinary English text, there are on an average about 4.5 letters per word [28]. English words form with only 5 vowels and 21 consonants whereas Bangla words form with 11 vowels, 20 allographs, and 39 consonants making the Bangla word length longer. The most frequent words are the function words of any language. Bangla function words are very short. If we didn't consider function words to calculate average word length then the average length per word would be even longer than 5.34 letters per word. In the article [18], authors found an average Bangla word length of 5.15 characters per word, and in [28] authors reported an

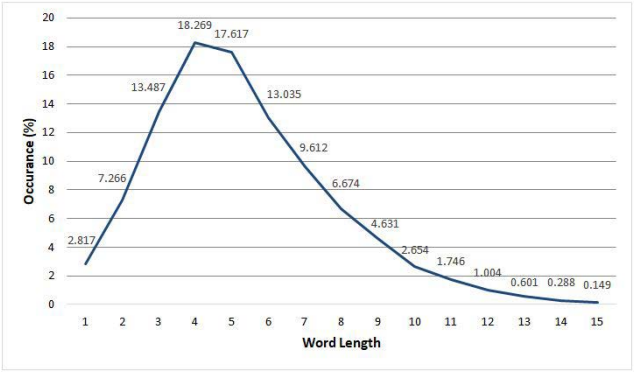

FIGURE 3. Usage of words vs. word length in KUMono corpus.

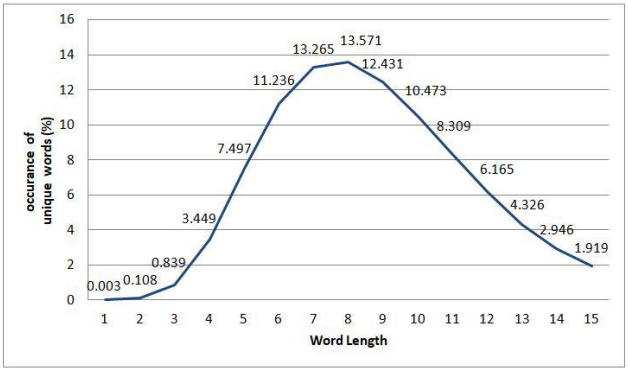

FIGURE 4. Usage of unique words vs. word length in KUMono corpus.

average Bangla word length of 5.12 for CIIL corpus which are similar to our findings. Whereas in article [15], authors found an average word length of 8.62 characters per Bangla word which is dissimilar to our result. Authors of [15] did not consider hyphens as word boundary but we did. Probably this is the reason for them behind finding a higher value for average word length. Figure 3 presents usage of words (%) vs. word length of our developed corpus. The high frequency of function words has an impact on Figure 3. We also experimented on % of occurrence of n-letter words vs. word length considering unique words only. Figure 4 presents the results for unique words. Though average word length of the corpus is 5.34, considering only % of occurrence of unique words, average word length is around 8. We also sort out frequent n-letter words of the corpus. Table 7, presents top 10 frequent n-letter words of our corpus. Words with joint letters count one extra letter because a Hoshonto (্) is counted with the first letter of the joint letter.

B. CHARACTER LEVEL ANALYSIS

1) FREQUENCY OF CHARACTERS

We have calculated the occurrence of each Bangla character in percentage according to our corpus. We found া, র, ে, ি, ি, ন, ক, ব are the most frequent characters of Bangla and ঔ, ঢ়, ঊ, ী, ঋ, ঐ, ঈ are the less frequent characters of Bangla. Table 8, presents the usage of each Bangla character in percentage according to our corpus. Character level statistical analysis of corpus shows that around 34.26% characters

TABLE 7. The top 10 frequent n-letter (n = 1 to 7) words in the KUMono corpus.

1-letter	%	2-letter	%	3-letter	%	4-letter	%	5-letter	%	6-letter	%	7-letter	%
ও	1.15	না	0.55	করে	0.84	থেকে	0.67	হয়েছে	0.55	কিন্তু	0.18	পর্যন্ত	0.16
এ	0.66	এই	0.51	করা	0.59	তিনি	0.50	মধ্যে	0.27	করেছেন	0.17	বিভিন্ন	0.15
র	0.10	হয়	0.50	এবং	0.46	বলেন	0.48	সঙ্গে	0.26	আমাদের	0.15	মাধ্যমে	0.10
ড	0.06	এক	0.33	হবে	0.42	জন্য	0.37	তাদের	0.25	হিসেবে	0.14	উপজেলার	0.09
ক	0.03	এর	0.26	তার	0.40	নিয়ে	0.34	হাজার	0.25	প্রধান	0.10	সম্পাদক	0.08
ই	0.02	আর	0.22	বলে	0.24	করেন	0.31	জানান	0.22	সাধারণ	0.09	মানুষের	0.07
খ	0.02	পর	0.21	হয়ে	0.24	করতে	0.29	হচ্ছে	0.19	আওয়ামী	0.09	সরকারের	0.07
আ	0.01	জন	0.21	সময়	0.20	একটি	0.28	করেছে	0.18	উপজেলা	0.08	নিশ্চিত	0.06
ঘ	0.01	গত	0.20	তবে	0.19	তারা	0.20	পুলিশ	0.16	সভাপতি	0.08	গ্রামের	0.06
গ	0.01	যে	0.19	কথা	0.18	দিয়ে	0.18	দেশের	0.14	অভিযোগ	0.07	স্থানীয়	0.05

TABLE 8. Percentage of occurrence of each letter in the corpus.

letter	%	letter	%	letter	%	letter	%	letter	%
া	10.578	য়	2.201	আ	1.002	ণ	0.428	ৎ	0.066
র	8.726	দ	2.176	ী	0.993	ফ	0.397	ঢ	0.047
ে	8.637	ু	1.772	চ	0.874	ড়	0.326	ঝ	0.041
্	6.395	য	1.718	ও	0.748	ঠ	0.188	ঃ	0.014
ি	5.767	হ	1.539	থ	0.712	ৃ	0.186	ঈ	0.012
ন	5.517	জ	1.522	ষ	0.700	ঙ	0.181	ঐ	0.009
ক	4.509	ট	1.454	ধ	0.692	ঘ	0.177	ঋ	0.007
ব	3.768	শ	1.356	ভ	0.680	ূ	0.164	ী	0.003
ত	3.529	ো	1.303	অ	0.589	ঁ	0.145	ঊ	0.002
স	3.234	গ	1.177	খ	0.573	ঁ	0.129	ঢ়	0.001
ম	2.967	এ	1.077	ং	0.507	ঞ	0.104	ঔ	0.001
ল	2.857	ছ	1.061	উ	0.439	ৌ	0.072		
প	2.429	ই	1.011	ড	0.436	ৈ	0.071		

of the corpus are vowels and around 65.64% characters of the corpus are consonants.

2) FREQUENCY OF INITIAL CHARACTERS

The frequency of each Bangla alphabet that starts a Bangla word in a Bangla article has been observed for the developed KUMono corpus. We have calculated the frequency of words with a certain initial character in two different ways. Firstly from total word tokens of the corpus and secondly from the unique unigrams of the corpus. In the total tokens list, words starting with ক has been used most frequently. 9.07% of words of the corpus starts with ক. ক, স, ব, প, ম have been used more frequently than other characters as initial characters of words. Table 9, presents a complete list of the percentage of occurrence of each alphabet as the initial character of Bangla word according to our corpus. We have observed that the probability of occurrence as first alphabet for ক, স, ব, প, ম are higher than the other ones. So, we can predict that we will get more words in a Bangla document starting with these five alphabets. Comparatively, ঊ, ঔ, ণ, ঙ, ঞ have less importance as starting alphabets of Bangla words. There are total 50 different Bangla alphabets (Bangla vowels and Bangla consonants). Few of them never occurs at the starting position of a Bangla word such as য়, ড়, ঢ়, ৎ, ং, ঃ, ঁ, these seven letters are never used as initial character of any word. We have also calculated the frequency of words with a certain initial character from the list of unique word tokens of the corpus. This means, how many unique unigrams the corpus have that starts with অ, আ, ক, খ etc. We have shown a list of initial letter frequency of Bangla alphabets from unique tokens in Table 10. The unique word list of the corpus contains more words with initial character স. There are 1,54,235 unique unigrams those start with alphabet স. So it has been observed from statistical data that though the corpus contains more unique words starting with স, yet the frequency of words starting with ক is higher than the frequency of words starting with স in the corpus.

In character level statistical analysis of KUMono corpus, we have done some other simple frequency based statistics. Among allographs the occurrence of aa-kar (' া ' কার) is highest (35.15%) followed by e-kar (' ি 'কার) and i-kar (' ী কার) respectively. Among consonant graphic variants ra-phala (র ফলা) (40.82%) is highest followed by j-phala (য ফলা) and b-phala (ব ফলা).

VI. QUALITY ASSURANCE OF KUMono CORPUS

It is unrealistic to ensure quality of a multi-million word corpus by manual checking. Thus, we evaluated the quality of the developed corpus in two ways. Firstly, we perform different empirical study based evaluation to assure that the developed corpus follows general features of natural language texts. Secondly, we implement article categorization as an application of the developed corpus and compare the results with state-of-the art categorization results to ensure quality of the corpus as training dataset for probabilistic models. Behaviour of function words are analysed to measure the homogeneity of the corpus, Zipf's curve is plotted with word frequencies and ranks to analyze whether the term distributions of the corpus data follows human nature of vocabulary usage, inherent

TABLE 9. Percentage of occurrence of initial letter of words in the corpus.

letter	%	letter	%	letter	%	letter	%	letter	%
ক	9.070	দ	4.244	য	1.859	ট	0.943	ঋ	0.037
স	8.927	ত	3.565	উ	1.579	ঘ	0.683	ষ	0.022
ব	8.678	জ	3.489	ভ	1.430	ড	0.672	ঊ	0.010
প	7.592	অ	3.192	থ	1.317	ধ	0.538	ঔ	0.004
ম	5.477	র	2.616	ল	1.149	ঢ	0.248	ণ	0.001
এ	5.294	শ	2.412	ই	1.082	ঝ	0.095	ঙ	0.001
আ	5.247	গ	2.101	ফ	1.019	ঠ	0.084	ঞ	0.001
হ	4.987	চ	1.918	খ	0.963	ঈ	0.048		
ন	4.527	ও	1.881	ছ	0.948	ঐ	0.046		

TABLE 10. Percentage of occurrence of initial letter of unique words in the corpus.

letter	%	letter	%	letter	%	letter	%	letter	%
স	9.148	জ	3.177	ত	2.059	ছ	0.787	ঋ	0.068
ব	8.439	গ	3.087	ল	2.025	ধ	0.695	ঐ	0.067
প	7.498	হ	3.046	ফ	1.977	ঘ	0.546	ঊ	0.056
ক	7.386	দ	2.808	ট	1.624	থ	0.397	ঔ	0.039
ম	7.268	শ	2.799	উ	1.618	ঝ	0.347	ঞ	0.018
আ	4.487	এ	2.778	ড	1.604	ঢ	0.144	ঙ	0.013
ন	3.838	চ	2.368	খ	1.189	ঠ	0.133	ণ	0.011
অ	3.805	ই	2.187	ও	1.185	ঈ	0.097		
র	3.586	ভ	2.129	য	0.848	ষ	0.089		

sparseness of the developed corpus is analyzed and compared with similar sized Arabic and English corpus, vocabulary growth rate of the corpus is measured for understanding how frequently new word types are encountered.

A. ZIPF's LAW

If all the words of a large corpus are listed in order of their frequency of occurrence then Zipf's law states that there is a relationship between the frequency of a word f and its rank in the list r [29]. According to Zipf's law:

$$f \propto \frac{1}{r} \tag{1}$$

There is a constant k such that $\boldsymbol{f}.\boldsymbol{r} = k$.

For example, this says that the 100th most common word should occur with three times the frequency of the 300th most common word. As the values of f and r tends to be very big, Zipf's law is rewritten as:

$$log(f).log(r) = log(k). \tag{2}$$

Figure 5 presents the Zipf's curve of our corpus. From Figure 5, we observe that the curve is almost linear for our corpus, so Zipf's law is approximately held. Zipf's law is useful as a rough description of the frequency distribution of words in human languages. There are a few very common words, a middling number of medium frequency words, and many low-frequency words. Human nature tends to minimize effort by using a small vocabulary of common words [29].

B. HAPAX LEGOMENA AND VOCABULARY GROWTH

Statistical results of the distribution of word types and their frequencies are given in Table 11. Close to half of the words

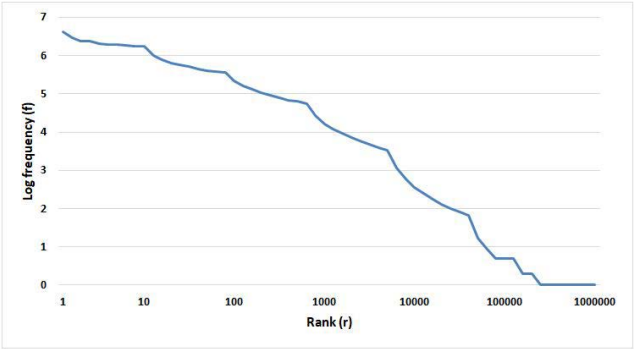

FIGURE 5. Zipf's curve of KUMono corpus.

are *hapax legomena*, i.e. occur only once. This is also true for most of the large corpora [3]. About 85 percent of the words are instanced just a few times. The vocabulary growth rate of any corpus can be calculated from *V(1)/N* (the ratio of the number of words that occurred once and the total number of words of that corpus) [30]. The written section of the British National Corpus has a vocabulary growth rate 0.003 and the vocabulary growth rate of SUMono corpus is 0.01 [18]. The vocabulary growth rate of our developed KUMono corpus is (731,912/353,547,583) = 0.0021 which is relatively lower than compared with SUMono corpus. As our developed corpus size is much larger than SUMono corpus both in total word and the total word appeared once context, its vocabulary is growing at a slower pace. In Figure 6 the vocabulary growth rate of KUMono corpus is depicted.

C. INHERENT SPARSENESS OF KUMono BANGLA CORPUS

Sparseness is an important issue in statistical natural language processing. Generally sparseness means almost all words in

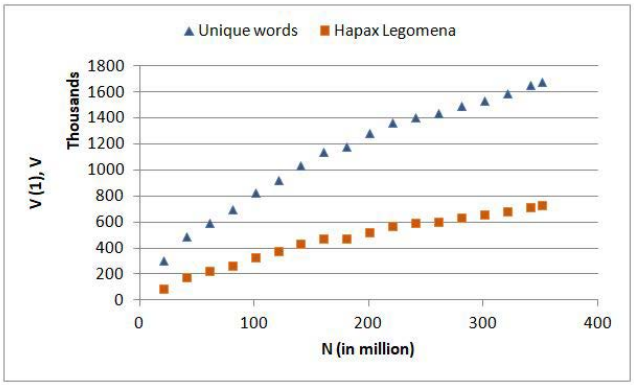

FIGURE 6. Vocabulary growth rate of KUMono corpus.

TABLE 11. Type token information of KUMono.

Words Frequency Count	No. of Words
Total number of word types:	16,86,270
Word-types occurring once:	7,31,912
Word-types occurring twice:	2,81,705
Word-types occurring 3-50 times:	5,55,875
Word-types occurring 51-100 times:	36,233
Word-types occurring 101-1000 times:	58,310
Word-types occurring 1001-5000 times:	14,455
Word-types occurring 5001-10,000 times:	3,131
Word-types occurring more than 10000 times:	4,649

a corpus are infrequent [29]. In other words, it means that some quite common or reasonable words are absent from a particular dataset. Inherent sparseness of a particular language is measured by calculating the Type-to-Token ratio of the language. For better understanding of inherent sparseness of any language, Type-to-Token Ratios of several languages are compared for equal text length and comparable genres. To study Type-to-Token ratio of the Bangla language in relation to English and Arabic according to our corpus, we have used Type-to-Token ratios of renowned Brown corpus and Al-Hayat corpus for English and Arabic language respectively. Information of Brown corpus and Al-Hayat corpus was taken from the article [19]. It may take more words in one particular language to express the same meaning than in another language. We present Type-to-Token ratio of these three languages for six different fragment sizes of data as displayed in Table 12. From Table 12 we observe that Bangla has a low Type-to-Token ratio than English and a high value than Arabic for all fragment sizes. Morphosyntactic features and orthographic conventions influence Type-to-Token ratios for the same fragment size [19]. Arabic has comparatively rich vocabulary and morphological variation than English and Bangla. It means that any specific word will appear less frequently in Arabic than in English or Bangla. As a result, Arabic datasets will have a higher degree of inherent sparseness than comparable other language counterparts. Bangla is also rich in vocabulary and has high morphological inflection than English and is inherently more sparse than English. Sparseness also partially depends on corpus quality and corpus size. Sparseness becomes less acute as the corpus gets larger [31].

TABLE 12. Type-to-token ratios for corpora fragments of different lengths of different language.

Text Length	Bangla (KUMono)	English (Brown)	Arabic (Al-Hayat)
100	1.255	1.449	1.190
1600	1.809	2.576	1.774
6400	2.561	4.702	2.357
16000	3.378	5.928	2.771
20000	3.668	6.341	2.875
1000000	17.642	20.408	8.252

TABLE 13. Type-to-token ratio comparison of different Bangla corpora.

Text Length	KUMono (our developed)	SUMono [18]	Prothom-Alo [15]
100	1.255	1.204	1.136
1600	1.809	1.913	1.984
6400	2.561	2.455	2.385
16000	3.378	2.985	3.135
20000	3.668	3.244	3.366
1000000	17.642	15.859	14.855

Each fragment's Type-to-Token ratio presented in Table 12 is average value of five random fragments of respective sizes. In Table 13 we present a Type-to-Token ratio comparison of Prothom-Alo [15], SUMono [18], and our developed KUMono corpus. It is evident that for the same fragment size our developed KUMono corpus is less sparse than the other two comparative corpora.

D. BEHAVIOUR OF FUNCTION WORDS

We have removed the function words from one version of our corpus and kept them in another version to analyze their influence on language formation and the behavior of function words of the Bangla language. The function words are named as ''bursty'' words as they occur more frequently in all types of documents. In our corpus, 35.57% of tokens of the whole corpus are function words or stop words.

Though function words have little value in information retrieval or knowledge engineering, their homogeneous presence across the corpus proves corpus quality. To analyze the behavior of function words in our corpus, we randomly took five equal-sized chunks from the corpus and calculated the word frequencies. In Table 14, top frequent 10 words from each chunk are presented along with their percentage of the occurrence. Because function words tend to distribute more homogeneously than content words of any corpus, their percentage of occurrence should have a similar value or they should appear in a similar order in all comparable chunks for equal-sized chunks of any corpus. In Table 14, function words occurring in similar order and their percentage of occurrence in the different chunks are also similar. So the function words are distributed homogeneously across the corpus.

VII. ARTICLE CLASSIFICATION USING THE KUMono CORPUS

We have implemented automatic article categorization as an application of our developed corpus. With the rapid growth

TABLE 14. Behaviour of function words.

Chunk 1		Chunk 2		Chunk 3		Chunk 4		Chunk 5	
Word	% of occurrence	Word	% of occurrence	Word	% of occurrence	Word	% of occurrence	Word	% of occurrence
ও	1.26	ও	1.21	ও	1.23	ও	0.93	ও	1.24
করে	0.99	করে	0.82	করে	0.76	করে	0.90	করে	0.83
এ	0.96	এ	0.73	এ	00.74	থেকে	0.68	এ	0.74
থেকে	0.75	থেকে	0.72	থেকে	0.72	হয়	0.63	থেকে	0.68
করা	0.72	করা	0.63	না	0.63	এ	0.60	করা	0.61
হয়েছে	0.68	হয়েছে	0.60	হয়েছে	0.65	করা	0.56	হয়	0.75
হয়	0.60	না	0.57	হয়	0.57	হয়েছে	0.54	হয়েছে	0.70
তার	0.57	হয়	0.55	এই	0.47	এবং	0.51	তিনি	0.66

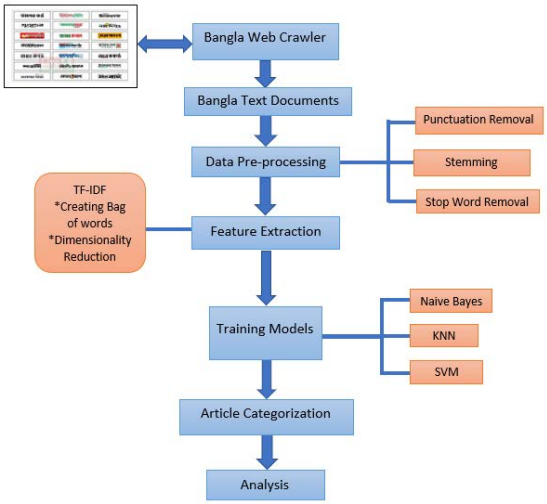

FIGURE 7. System flow diagram of the developed article categorization model.

and availability of digital texts, automatic categorization of texts can be useful for organizing huge digital content. The process of labeling articles into one of the predefined categories based on content is article categorization. The developed corpus has a total of 1,292,527 articles. Instead of taking the whole corpus, we have taken 8 categories and 30K articles from each category of the developed KUMono corpus for categorization.

A. ARTICLE CATEGORIZATION MODEL

The steps of classification include data selection from the KUMono corpus, pre-processing, feature extraction, and classifier training. We categorize news articles based on the result of the trained model and finally analyze our results. Figure 7 depicts the total article categorization model. Here pre-processing steps include tokenization, punctuation removal, stop word removal, and stemming. Data collection and pre-processing steps are already discussed in the previous section IV. Rest of the steps are discussed in the following sub-sections.

1) FEATURE EXTRACTION

We have used the well-known feature extraction metric TF-IDF (Term Frequency-Inverse Document Frequency), to extract features from articles. TF-IDF is a weighting scheme often used in information retrieval and text mining [32]. It is a numerical statistic that is intended to reflect how important a word is to a document in a corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word. The TF-IDF value of all the words has been calculated to have an insight into the total corpus.

2) CLASSIFIER DESIGN FOR TRAINING

The classifiers we have adopted to perform article categorization are: Naïve Bayes (NB), Support Vector Machine (SVM), and K Nearest Neighbor (KNN), Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF).

- *Naïve Bayes (NB):* Naïve Bayes (NB) is a simple probabilistic classifier based on Bayes' theorem [33]. It is a supervised learning algorithm and can make quick predictions by building fast machine learning models. It can be used for binary and multi-class classifications.
- *Support Vector Machine (SVM):* SVM is a supervised machine learning technique which is actually a binary classifier based on the principle of structural risk minimization [34]. SVM creates a hyperplane or decision boundary to segregate n-dimensional space into classes.
- *K Nearest Neighbor (KNN):* KNN is a statistical pattern recognition algorithm that has been studied extensively for text categorization applications [35]. It is a simple supervised learning technique that classifies new data into a category based on similarities. KNN algorithm measures similarities between new data and available previous cases and put the new case into a category that is most similar to the data.
- *Logistic Regression (LR):* Logistic Regression is a supervised classification algorithm. It fits a logistic function also called sigmoid function to predict probabilistic values of depended variables. The curve generated from the logistic function indicates the likelihood of occurrence of something.
- *Decision Tree (DT):* Decision Tree is another supervised machine learning classifier that produces a tree shaped structure where the internal nodes represent dataset features, branches represent decision rules and the leaf nodes represent the outcomes.
- *Random Forest (RF):* Random Forest is an ensemble learning based machine learning technique. It forms multiple decision trees out of the dataset and predicts the final output based on the majority votes of the

TABLE 15. Performance of decision tree.

Category	Precision	Recall	F1-score
Crime	1.00	1.00	1.00
Economics	0.98	0.98	0.98
Education	0.98	0.98	0.98
Entertainment	0.95	0.98	0.97
Politics	0.99	0.98	0.98
Religion	1.00	0.99	0.99
Sports	0.97	0.95	0.96
Technology	0.96	0.96	0.96
Macro Avg.			**0.98**
Accuracy			**0.98**

TABLE 16. Performance comparison of different learning models.

Learning Model	Precision	Recall	F_1-score
Naïve Bayes	0.81	0.80	0.80
SVM	0.80	0.79	0.79
KNN	0.79	0.79	0.79
Logistic Regression	0.78	0.77	0.77
Random Forest	0.98	0.98	0.98
Decesion Tree	0.98	0.98	0.98

predictions of the decision trees. The greater the number of decision trees in the forest, the better the prediction with increased computational cost.

3) PERFORMANCE METRIC

To evaluate the performance of the developed article categorization method we have used the F_1-measure and accuracy. F_1 measure is directly related to precision and recall. Precision is the ratio between the true Positives and total predicted positives. It reveals how accurate the model is compared to all predicted positives. And Recall is the ratio between the true positives and total actual positives. So recall reflects how many positives the developed model can detect compared to all actual positives. Precision and recall are calculated as the following equations:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Here TP, TN, FP and FN represents True Positive, True Negative, False Positive and False Negative respectively. F_1-score is calculated from precision and recall by the following equation,

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

F_1-measure is the harmonic mean of the model's Precision and Recall. Accuracy is the ratio between the number of correct predictions and total predictions. The best accuracy value is 1 and 0 is the worst. It reflects how accurate a model is behaving. Accuracy can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

B. EXPERIMENTAL RESULTS AND DISCUSSIONS

We have developed an experimental dataset consisting of articles from the corpus. In article categorization dataset, articles belonged to the following categories: Crime, Economics, Entertainment, Politics, Religion, Sports, Technology, and Education. The dataset consists of 30,000 articles from each category. The training and test data consists of 90% and 10% data from total dataset. We found the best results for Random Forest (RF) and Decision Tree (DT) algorithms. In Table 15 the details Precision, Recall, F1-score and Accuracy of each category is presented for Decision Tree (DT) classifier.

In Table 16 the average precision, recall, F_1-score of each classifier is presented. Here the values presented are the average value of all categories. From Table 16 we can easily identify that Random Forest and Decision Tree outperformed other classifiers. It is noted that TF-IDF feature extraction method was applied for all classifiers.

In Table 17 we compared the average performance of Random Forest classifier on our article categorization experimental dataset with the best results of other state-of-the-art Bangla document classification research. From Table 17 it is clear that our classification result outperformed all the state-of-the-art works except [36]. Deep learning based model LSTM proposed in [36] slightly outperformed our model. Better performance of classification largely relies on the type of learning model and quality of the training dataset. Deep learning based learning models are able to extract the

TABLE 17. Performance comparison with state-of-the-art works in Bangla text classification.

Learning Model	Precision	Recall	F_1-score	Accuracy(%)
Random Forest (On proposed KUMono dataset)	**0.98**	**0.98**	**0.98**	**98%**
BiLSTM [36]	0.98	0.98	0.98	98.33%
Word2Vec (NN) [26]	0.96	0.96	0.96	96%
TF-IDF* (NN) [26]	0.96	0.96	0.96	96%
TF-IDF (SGD) [37]	0.92	0.92	0.92	92%
TF-IDF (SVM) [37]	0.92	0.92	0.92	92%
SVM [38]	N/A	N/A	0.89	89.14%
Naive Bayes [38]	N/A	N/A	0.85	85.22%
DenseNN2 [39]	0.85	0.85	0.85	85.20%

contextual information of data by adding extra computational cost. We believe, machine learning algorithms Random Forest and Decision Tree performs almost same as the best result of state-of-the-art research because of the quality of training dataset.

VIII. CONCLUSION

Bangla language is lacking in enough NLP research and resources compared to many other languages. In this article, we have addressed the issue of resource building by developing a monolingual Bangla corpus of more than 353 million word tokens. We also have performed a detailed analysis of several word level and character level phenomena of the Bangla language by various statistical tools. We have studied the usage frequency of words of different lengths, the average word length of Bangla, the most frequent words in Bangla, etc. Character level analysis of Bangla language includes observing the most frequent and least frequent characters, characters frequently used as word initials, frequency of unique tokens of specific initial characters, etc. Assessment of the corpus is done by applying Zipf's law on the corpus, assessing the homogeneity of distribution of function words in Bangla, and analyzing the inherent sparseness of Bangla by comparing the type-token ratio of Bangla with English and Arabic. We found that our developed corpus roughly follows Zipf's curve. Like other large corpora, it also follows *hapax legomena* characteristics. We have discussed article categorization as an application of the developed corpus and our categorization results outperformed other state-of-the-art Bangla document categorization results.

Other possible applications could be N-Gram-based spell checking and correction tool design. We can also build an English transliteration dictionary of Bangla. The Association of this list with the dictionary can be used to analyze social media texts. Furthermore, the dictionary can be used to design a keywords-based search engine where articles are indexed by category. A good quality question answering system can be designed using the KUMono corpus. Also, article summarization and fake news detection or generation applications can be designed using this resource. In the future, we envision developing various applications based on the developed KUMono corpus. We hope NLP and linguistic researchers of Bangla would find this corpus useful in different aspects of research on the Bangla language.

ACKNOWLEDGMENT

The authors would like to thank the students of the Computer Science and Engineering Discipline, Khulna University, associated with the data collection process.

REFERENCES

[1] Wikipedia Contributors. *Bengali Language*. Accessed: May 20, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Bengali_language

[2] A. O'Keeffe and M. McCarthy, *The Routledge Handbook of Corpus Linguistics*. Evanston, IL, USA: Routledge, 2010.

[3] H. Steger, G. Ungeheuer, and H. E. Wiegand, *Handbücher zur Sprach-und Kommunikationswissenschaft*. Berlin, Germany: Walter de Gruyter, 1991.

[4] S. Greenbaum and J. Svartvik. *The London-Lund Corpus of Spoken English*. Accessed: May 22, 2021. [Online]. Available: http://korpus.uib.no/icame/manuals/LONDLUND/INDEX.HTM

[5] B. Megyesi, J. Näsman, and A. Palmér, "The Uppsala corpus of student writings: Corpus creation, annotation, and analysis," in *Proc. 10th Int. Conf. Lang. Resour. Eval. (LREC)*, 2016, pp. 3192–3199.

[6] University of Nottingham. *Cambridge and Nottingham Corpus of Discourse in English*. Accessed: May 20, 2021.[Online]. Available: https://www.nottingham.ac.uk/research/groups/cral/projects/cancode.aspx

[7] The University of Michigan English Language Institute. *Michigan Corpus of Academic Spoken English*. Accessed: Feb. 7, 2022. [Online]. Available: https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple

[8] G. Kennedy, *An Introduction to Corpus Linguistics*. Evanston, IL, USA: Routledge, 2014.

[9] S. Conrad, "Will corpus linguistics revolutionize grammar teaching in the 21st century?" *TESOL Quart.*, vol. 34, no. 3, pp. 548–560, 2000.

[10] T. McEnery and R. Xiao, "What corpora can offer in language teaching and learning," in *Handbook of Research in Second Language Teaching and Learning*, vol. 2. London, U.K.: Routledge, 2011, pp. 364–380.

[11] S. Koeva, I. Stoyanova, S. Leseva, R. Dekova, T. Dimitrova, and E. Tarpomanova, "The Bulgarian national corpus: Theory and practice in corpus design," *J. Lang. Model.*, vol. 1, no. 1, pp. 65–110, Dec. 2012.

[12] Cambridge University Press. *Cambridge English Corpus*. Accessed: May 2021. [Online]. Available: https://www.cambridge.es/en/about-us/cambridge-english-corpus

[13] Sketch Engine. *Oxford English Corpus*. Accessed: May 2021. [Online]. Available: https://www.sketchengine.eu/oxford-english-corpus/

[14] P. Baker, A. Hardie, T. McEnery, R. Xiao, K. Bontcheva, H. Cunningham, R. Gaizauskas, O. Hamza, D. Maynard, V. Tablan, C. Ursu, B. D. Jayaram, and M. Leisher, "Corpus linguistics and South Asian languages: Corpus creation and tool development," *Literary Linguistic Comput.*, vol. 19, no. 4, pp. 509–524, Nov. 2004.

[15] K. M. Majumder and Y. Arafat, "Analysis of and observations from a Bangla news corpus," BRAC Univ., Dhaka, Bangladesh, Tech. Rep., 2006.

[16] A. I. Sarkar, D. S. H. Pavel, and M. Khan, "Automatic Bangla corpus creation," BRAC Univ., Dhaka, Bangladesh, Tech. Rep., 2007.

[17] K. M. Anwarus Salam, M. Rahman, and M. M. S. Khan, "Developing the Bangladeshi national corpus—A balanced and representative Bangla corpus," in *Proc. Int. Conf. Sustain. Technol. Ind. 4.0 (STI)*, Dec. 2019, pp. 1–6.

[18] M. A. A. Mumin, A. Awal, M. Shoeb, M. R. Selim, and M. Z. Iqbal, "Sumono: A representative modern Bengali corpus," *SUST J. Sci. Technol.*, vol. 21, no. 1, pp. 78–86, 2014.

[19] A. Sarkar, A. De Roeck, and P. Garthwaite, "Easy measures for evaluating non-English corpora for language engineering: Some lessons from Arabic and Bengali," Dept. Comput., Fac. Math. Comp., Open Univ., Walton Hall, U.K., Tech. Rep. 2004/05, 2004, pp. 1–5.

[20] N. Hossain, S. Islam, and M. N. Huda, "Development of Bangla spell and grammar checkers: Resource creation and evaluation," *IEEE Access*, vol. 9, pp. 141079–141097, 2021.

[21] K. M. Alam, M. T. H. Hemel, S. M. M. Islam, and A. Akther, "Bangla news trend observation using LDA based topic modeling," in *Proc. 23rd Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2020, pp. 1–6.

[22] A. Amin, I. Hossain, A. Akther, and K. M. Alam, "Bengali VADER: A sentiment analysis approach using modified VADER," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2019, pp. 1–6.

[23] S. Islam, M. A. Lotif, A. Akther, and K. M. Alam, "BNNet: Towards a comprehensive Bangla WordNet," in *Proc. 23rd Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2020, pp. 1–6.

[24] N. J. Prottasha, A. A. Sami, M. Kowsher, S. A. Murad, A. K. Bairagi, M. Masud, and M. Baz, "Transfer learning for sentiment analysis using BERT based supervised fine-tuning," *Sensors*, vol. 22, no. 11, p. 4157, May 2022.

[25] N. S. Dash, B. B. Chaudhuri, P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja, "Corpus-based empirical analysis of form, function and frequency of characters used in Bangla," in *Proc. Corpus Linguistics Conf.*, vol. 13, P. Rayson, A. Wilson, T. McEnery, A. Hardie, and S. Khoja, Eds. Lancaster, U.K.: Lancaster Univ. Press, 2001, pp. 144–157.

[26] M. T. Alam and M. M. Islam, "BARD: Bangla article classification using a new comprehensive dataset," in *Proc. Int. Conf. Bangla Speech Lang. Process. (ICBSLP)*, Sep. 2018, pp. 1–5.

[27] D. M. F. Khan, A. Ferdousi, and M. A. Sobhan, "Creation and analysis of a new Bangla text corpus BDNC01," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 5, no. 11, pp. 260–266, Nov. 2017.

[28] N. S. Dash, *Corpus Linguistics and Language Technology: With Reference to Indian Languages*. New Delhi, India: Mittal Publications, 2005.
[29] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.
[30] R. Baayen, *Word Frequency Distributions*. Dordrecht, The Netherlands: Kluwer, 2001.
[31] A. Goweder and A. De Roeck, "Assessment of a significant Arabic corpus," in *Proc. Arabic NLP Workshop ACL/EACL*, 2001, pp. 73–79.
[32] A. Dhar, N. S. Dash, and K. Roy, "Categorization of Bangla web text documents based on TF-IDF-ICF text analysis scheme," in *Proc. Annu. Conv. Comput. Soc. India*. New York, NY, USA: Springer, 2018, pp. 477–484.
[33] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5432–5435, Apr. 2009.
[34] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Apr. 1995.
[35] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1999, pp. 42–49.
[36] S. Rahman and P. Chakraborty, "Bangla document classification using deep recurrent neural network with BiLSTM," in *Proc. Int. Conf. Mach. Intell. Data Sci. Appl.* New York, NY, USA: Springer, 2021, pp. 507–519.
[37] M. S. Islam, F. E. M. Jubayer, and S. I. Ahmed, "A comparative study on different types of approaches to Bengali document categorization," 2017, *arXiv:1701.08694*.
[38] A. K. Mandal and R. Sen, "Supervised learning methods for Bangla web document categorization," 2014, *arXiv:1410.2045*.
[39] M. Chakraborty and M. N. Huda, "Bangla document categorisation using multilayer dense neural network with TF-IDF," in *Proc. 1st Int. Conf. Adv. Sci., Eng. Robot. Technol. (ICASERT)*, May 2019, pp. 1–4.

AYSHA AKTHER received the B.Eng. degree from Khulna University, Bangladesh, and the M.C.S. degree from the University of Ottawa, Canada. She is currently pursuing the Ph.D. degree in computer science and engineering with Khulna University.

She was awarded the University of Ottawa Graduate Scholarship during her master's studies. She has authored and coauthored 13 peer-reviewed international conference and journal articles with the ACM, IEEE, and Springer publishers. Her research interests include natural language processing, machine learning applications, and data mining applications.

MD. SHYMON ISLAM received the B.Eng. degree in computer science and engineering from Khulna University, Bangladesh, in 2022, where he is currently pursuing the M.Eng. degree.

His research interests include natural language processing, machine learning, data mining, pattern recognition, image processing, design and analysis of algorithm, and metaheuristic. He has authored one peer-reviewed international conference with the Springer publishers and received best paper award in the conference.

HAFSA SULTANA received the B.Eng. degree from Khulna University, Bangladesh, in 2022, where she is currently pursuing the M.Eng. degree.

Her research interests include natural language processing, machine learning, artificial intelligence, and deep learning.

A. K. Z. RASEL RAHMAN received the B.Eng. degree from Khulna University, Bangladesh, in 2022, where he is currently pursuing the M.Eng. degree.

His research interests include algorithms, artificial intelligence, computer vision, data mining, machine learning, natural language processing, programming environments, and robotics.

SUJANA SAHA received the B.Eng. degree from Khulna University, Bangladesh, in 2022.

Her research interests include deep learning, natural language processing, and computer vision.

KAZI MASUDUL ALAM received the B.Eng. degree from Khulna University, Bangladesh, and the M.C.S. and Ph.D. degrees from the University of Ottawa, Canada.

He is currently working as a Professor of computer science and engineering with Khulna University. During his graduate studies, he played key roles in the design process of the social Internet of Vehicles, digital twin architecture, haptic e-book, and haptic EmoJacket. He has authored and coauthored 30 peer-reviewed international conference and journal articles with the ACM, IEEE, and Springer publishers. His research interests include blockchain, the IoT, smart city applications, machine learning, and natural language processing. He has received best paper awards in a few IEEE conferences. He was awarded prestigious academic and research scholarships of NSERC Canada Graduate for his doctoral degree and the Ontario Graduate Scholarship for his master's degree during his graduate studies. He regularly reviews journal articles from the top publishers of IEEE, ACM, and Springer as well as reviews project proposals for international funding bodies.

RAMESWAR DEBNATH received the bachelor's degree in computer science and engineering from Khulna University, Bangladesh, in 1997, and the Master of Engineering degree in communication and systems and the Ph.D. degree from The University of Electro-Communications, Tokyo, in 2002 and 2005, respectively, under the Japanese Government Scholarship.

He worked as a Postdoctoral Researcher under the JSPS Fellowship with the Department of Informatics, The University of Electro-Communications; the Neuroscience Research Institute; and the National Institute of Advanced Industrial Science and Technology, Tsukuba, from 2008 to 2010. Currently, he is working as a Professor with the Computer Science and Engineering Discipline, Khulna University. He has published more than 40 journals, book chapters, and peer-reviewed conference papers. He presented papers in many conferences in home and abroad. His research interests include artificial intelligence, machine learning, pattern recognition, image processing, and bioinformatics. He was the Organizing Chair of the 16th International Conference on Computer and Information Technology (ICCIT), in 2014.