Utilizing Optical Character Recognition to Summarize Bengali Text

Abhijit Saha, Faiza Bushra, Ahmed Anwar, Rubaba Rashid, Md Sabbir Hossain, Ehsanur Rahman Rhythm and Annajiat Alim Rasel

Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University

66 Mohakhali, Dhaka - 1212, Bangladesh

{abhijit.saha, faiza.bushra, ahmed.anwar, rubaba.rashid, md.sabbir.hossain1, ehsanur.rahman.rhythm}@g.bracu.ac.bd, annajiat@gmail.com

Abstract—Optical Character Recognition (OCR) has proven to be a valuable tool for extracting textual information from images, converting physical documents into digital formats that can be easily edited, searched, and shared. However, OCR technology is not without its limitations, particularly when it comes to recognizing and processing non-Latin scripts such as Bengali. Bengali is one of the most widely spoken languages in the world, with over 230 million speakers worldwide. One area where OCR could be particularly useful is in summarizing Bengali text, which could save time and effort for researchers, journalists who are working with a large corpus of text. In recent years, there have been several attempts to develop OCR-based text summarization techniques for Bengali. One challenge it struggles with is the complex nature of the language, its rich vocabulary, its highly inflected grammar and the inclusion of a number of ligatures in Bengali Script and diacritics that can significantly complicate the OCR process. Overall, the development of OCR-based text summarization could save time and effort, while also providing a more objective and systematic approach to analyzing large volumes of text. As the technology continues to improve, it is likely that we will see more sophisticated and accurate OCRbased text summarization techniques for Bengali in the future.

Index Terms—OCR, Bengali, Corpus, Summarization, Ligatures

I. Introduction

OCR, or Optical Character Recognition, was developed to automate the process of recognizing and converting printed or handwritten text into digital text that can be edited, searched, and stored electronically. Before the development of OCR technology, the process of digitizing text was time-consuming and labor-intensive. To convert printed or handwritten documents into digital text, individuals had to manually type out the text, a process known as manual data entry. This was not only time-consuming, but it was also prone to errors, which could lead to inaccuracies and lost productivity.

OCR technology was developed to automate the process of digitizing text, making it faster and more accurate. OCR software uses advanced algorithms to analyze the shapes and patterns of printed or handwritten characters and then recognizes and converts them into digital text. This enables users to easily convert large volumes of printed or handwritten

documents into digital format, saving time and improving accuracy. OCR technology eliminates the need for manual data entry by automating the process of converting scanned documents into machine-encoded text. This saves time and reduces errors that can occur during manual data entry, makes printed materials accessible to people with visual impairments by converting them into digital text that can be read aloud by screen readers and allowing documents to be scanned and translated into other languages using machine translation algorithms, saving time and improving accuracy.

As powerful as the technology can be, its limitations prevail when functioning non-latin languages. Bangla is the seventh most widely spoken language in the world and has a rich cultural heritage. Many important historical and literary works are written in Bangla. Developing Bangla OCR technology can help to preserve this cultural heritage by making it easier to digitize and store historical and literary works in digital format, and can help to make digital information more accessible to people who speak or read Bangla, further enhancing education by making it easier to share educational resources across different regions and countries.

However, the complex nature of Bengali script has made it difficult for OCR technology to accurately recognize and digitize Bengali text. This has been a major obstacle for researchers and others who need to work with Bengali texts, as it has made it difficult to search and analyze large volumes of Bengali text quickly and efficiently. Despite the development of Bangla OCR technology since the 1980s, the current OCR systems have not been able to achieve the desired level of accuracy. As a result, the need for improving Bangla OCR has become a major research area today. But with the resources available today, the implementation of a fully fledged Bangla OCR is very much feasible and its utilization can produce accurate outcomes along with reducing processing time.

II. LITERATURE REVIEW

One of most researchable topic in NLP is text summarization. OCR (Optical Character Recognition) technology can convert printed text or scanned images into machine-readable digital

text. Because this digitization automates the extraction of relevant information from documents, it reduces the requirement of manual summarization efforts and it is mostly used in text summarization process.

Many studies have been conducted in this field for various languages. To avoid ambiguity or contradiction, the process of calculating similarity via a basic pattern matching algorithm requires the possession of an exhaustive set of patterns for each meaning. The difficulty and time-consuming nature of manually compiling is tedious task. Recent natural language processing applications highlight the need for an efficient method to obtain the similarity between very textual data or sentences [1]. The utilization of a lexical dictionary to evaluate the similarity of a set of words collected from several sentences being compared is one extended form of word co-occurrence methods. Sentence similarity can be calculated simply by adding the similarity values of all word pairs[2]. There is also a model that uses a reinforcement technique for abstractive text summarization and employs convolutional sequence learning. The central concept of this work is textual learning in context. The convolutional neural network consolidates the entire process [3].

III. OCR APPROACHES

IV. SUMMARIZATION APPROACHES

V. DATA COLLECTION AND PREPROCESSING

A. Data Collection

For Summarization: To train a model for the purpose of summarization we collected around 23422 texts from the Prothom Alo Newspaper and their subsequent summarized texts. The texts were categorized into 7 types including-

- 1) Business News Data
- 2) Entertainment News Data
- 3) Opinion News Data
- 4) Politics News Data
- 5) Sports News Data
- 6) World News Data
- 7) Bangladesh News Data

Subsequent csv files were made for each type for data preprocessing. Each csv file had 6 columns- Title, Description, Meta-Summary, Summary, Syndicate-catagorys, Keywordsfor-related-articles. An illustration of which is shown below:

1 ber_test * ph.resf_test("/content/desiral/bjerses/SEADLTask-AA/Summeriantion_desiralses/assassassibles.co") 1 ber_test.head(15)						
	Title	Description	Reta-Summery	Summery	syndicate-catagorys	Keywords for related articles
0	रेका त्यवरक पे-सरका कृत विशेष van विश्वतिक विश्वता	%-বাবেন্তের মানিকপক বেন বাশিরা মন্ত্রণানরের	মুক্তরোগী প্রান্তবংক পক্ষে চরার্ক্তন আরু ফুলার সচিত্র	ভূমতেশী প্রসমবের পক্ষে চরামন আম কুনার সচিব	THU	रेक्सर समित
1	ট্রিংকরে মন্ত্রনিক্তে গুল্ম বৃদ্ধির প্রকাশ ভিয়ে	निरमणे देशीय स्थान करियान द्विश्वाय सञ्ज्ञीत	विद्यालगरमा क्षेत्र गाहारम परम द्वित्तमाता पर खाडान	विदारमध्या कह गहरूम यह द्वित्तारम यह स्थापन	n/has	कामसभि अकुमि, निरम्पी भाव
2	THE R CHICAL SCHOOL CARRESTON	भूगक के तमस्त्रापन करवाई त्यांच्या त्यांच्या स्थापन करवाई पूरण	When the contrast states of the contrast states of the α	When the contrast section of the contrast states of the $\chi_{\rm CC}$	viftos	fourt fort forter
3	अमीज स्मित समित्र सिमिन स्थानस	বংলাবেশ কৃষি বাংকের উপবাস্থাপন পরিচালক শ	বংলাদে কৃষি বাংকের উপরাবস্থাপন পরিচালক শ	যালালে কৃষি যানকর উপরবস্থাপন পরিলের শ	NaN	100
4	মালা পাঢ়ির অভেনারে মেচেন্দ	মল পার্বি, বাংলাবেশর পার্বির জন্মর প্রথম সুপরিতি	ारमा काराज्य कुल निव्हामधी कारमाता अस्पत ह	टारमा कवारूप कुल निवृत्त्वाची कारणवात अरुपा छ	NaN	देशका, शामादि, कारित
5	স্থাক প্ৰতিক্ৰমের স্থানীকলিক স্বাবহুর স্বাকলি মান	পালের মায়ুলর বিরস্তুর সেরহা প্রক মারেই গ্র	পাচৰাৰ ৰাষ্ট্ৰপা নিয়ন্ত্ৰণ দেৱা কৰা কাৰ্যেই হৈ	সংক্ৰমৰ মন্ত্ৰেলয় নিয়ন্ত্ৰণ কেন্তৰ প্ৰঞ্জ মন্ত্ৰকৌ হৈ	viftos	অফার্যবিক্রার, মানের, অম্বর্তিক
c	representa nera titore concensarios	राज्यभीत चारावर्गात असमात्र गीत अमारत गीते, था	शास्त्रभवित चालावर्णक अस्तावात गीत अक्तारत पत्ति, था	प्राप्तिक प्राप्तार्थक अगस्य गीव अस्तराव गीवे, थः	n/Hug	शक्सि, विवास क्षत्र, कारिति
,	যাপের চালকল ৬০০ কোটি টাকার কিবছে আকিছে প্রাপ	সম্ভাগনির ব্যোগায়টিকে ২০১৮ সালে প্রায় সেচু শ কোটি	ज्ञासम्बद्धाः राज्याचार्वरः २०५५ मारम द्वार राष्ट्र म रकार्ति	ज्ञासम्बद्धाः त्याचाराहीतम् २०५५ मारम क्षत्रः तस्य म दकारी	n/Has	्रक्तिकारमञ्जूषात् वरिका, वरिका गः वर
8	के सारशासन विकास 3,500 तमार्थि क्रिका साम्यागासक	*t-arcrif of bits in *t-occupy with a very con	रे-कार्या अविकास रे-कारका कुत्र सारिकालक असः	रे-कार्या श्रीवेशन रे-कार्यातुत प्रतिकशक श्रासः	tritus	र कावार् शर्मिका, शरावश
,	ছাতু বিয়ে এবার টিকে থাকার লড়বিরে নামহে পর্বাস থাক	পৰ্যান খাবের বারস্থানির টিকে খাববে চান। এ ফ্রন্ম	পর্যান থাকের ব্যবস্থানির টিকে থাকার চান। এ জনা	পর্যান থাকে কালভারির টিকে থাকার চন। এ জনা	n/has	পৰ্যান, বাৰসাধী, ৰয়োনা কৰ্মনিক
90	সমায়মকে ঋণ পোলে ১% প্রশোধনা পারেল নারী উধ্যোক্তারা	সময়মার মধ্যে দিবি পরিপার মধ্যে নারী ইচারকার	সময়মধ্যে ছবার কিন্তি পরিশেষ করলে মার্টি উচ্চাঞ্চার	সময়মধ্যে মধ্যে কিন্তি পরিশাস করলে মার্টি উলোক্তর	xiftus	वत्तराम सारम, यही, बन
11	NUMBER OF THE CONTRACT NUMBER COOPS NOT COT	राज्य क्रमावय कुल निवासकी कारमधार अस्पर है	ात्मत क्षमावय कृत निवृत्त्वती कारमाव अस्ता ह	ारभा क्षमाका कृत निवृत्त्वती क्षारमावा अस्ता ह	10100	कामानी, तराव, पुट्टा, निकार क्षतिरकात
2	सूनरे (बार मार्ग सारकारीया पात भवितास्था	পৰ বোৰনা বালেনৰ আৰুগনিবাদের মাঞ্চপনি ককুল	त्रव द्वारात शरमान शामगरिकारन सामगरि स्तूरा.	গৰ বোৰাৰ বাদেয়ৰ আৰক্ষমিকানে বাজনাৰী কাছুল	n/has	অফগনিয়ান, মাচি, কর্মগংহান
13	SIGON Yours quiles que sesses suféres	Salidan gar well wern segres cores m	Riflas over mende foliktion spare Riflass waces	Rika om nenik filiktim gan Rikas wers	effec	पुरा दीन पुरुषक्षे उत्तर प्रायप्
56	व्यक्तिकान निर्माण करते हमा दिविक	कट्टक मध्य प्राप्त कर रक्षा प्राप्तवाना प्राप्तवानी ।	करकर गर्भार प्रदेश कर रक्षा परस्थाना प्रकाशनिका	করেক সভাব আগে ভাল বঙ্গা বালেয়ানের আনগানিকা	160	WANTED THE WATER SPICES

Fig. 1. Example of a figure caption.

B. Data Preprocessing

For Summarization: We applied multifarious preprocessing techniques on the collected dataset in csv format to ensure that it was in a suitable format for training our model. For each of the 7 csv files we checked for duplicate and missing/null values. A few duplicate texts were found which might have been mistakenly taken during data collection. We removed the rows with duplicate values. Meanwhile for the null values we used an imputation technique. We got null values for the 'syndicate-catagorys' column of the csv file only. As the dataset was already divided, it was easy to replace the null values with its subsequent category using .fillna() function of the pandas library. Finally after preprocessing, all the refined datasets were combined together into 1 csv file to train our model.

VI. METHODOLOGY VII. RESULTS AND ANALYSIS VIII. CONCLUSION

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

CONCLUSION

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

- [1] [1] D. Michie, "Return of the Imitation Game,"
- [2] [2] Juan-Manuel Torres-Moreno, Automatic Text Summarization (Cognitive Science and Knowledge Management) 1st Edition, 2014.
- [3] [3] Gehring, Jonas, et al. "Convolutional sequence to sequence learning." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017

- [4] K. Elissa, "Title of paper if known," unpublished.
- [5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
- [6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
- [7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.