Closed Domain Bangla Extrinsic Monolingual Plagiarism Detection and Corpus Creation Approach

Adil Ahnaf, Shadhin Saha, Nahid Hossain
Department of Computer Science and Engineering
United International University, Dhaka, Bangladesh
aahnaf151054@bscse.uiu.ac.bd, ssaha151190@bscse.uiu.ac.bd, nahid@cse.uiu.ac.bd

***Abstract*— Plagiarism is an act of presenting other's works or ideas as someone's own. In recent days, online publishing of text through internet media is rapidly increasing and plagiarism is one of the most dangerous side effects of that. Plagiarism detection in academic institutions has become a regular feature in recent years. There are several approaches available for detecting plagiarism in the English language and other western languages. However, there is no plagiarism detection tool available that supports plagiarism detection in Bangla language where Bangla is one of the most spoken languages in the world and widely used on the internet. Therefore, we have developed a first-ever plagiarism detection tool for Bangla language. To reduce the domain complexity at this stage, we have reduced our domain to textbooks only. We have also constructed a corpus which consists of all books of the National Curriculum and Textbook Board (NCTB) of Bangladesh from class I-XII. Our proposed approach shows an accuracy of 96.75% at this stage.**

Keywords—Plagiarism, Bangla, Textbooks

I. INTRODUCTION

Plagiarism is a crime. Plagiarism is one kind of theft where presenting other's works or ideas as someone's own without full acknowledgement of source. Plagiarism exists in every field especially in journalism (i.e., newspapers, magazines, blogs) and academia (i.e., student reports, assignments, thesis). Most of the time entire document is exact copy otherwise modifications of single or multiple sources [1]. Plagiarism detection in academic institutions has become a regular feature in recent years. According to a study on 18,000 students shows that about 50% of the students admitted they usually plagiarize their theses and assignments from extraneous documents [2]. Academic institutions all over the world try to infix ethical values among students. As a result, the plagiarism detection tool is very important. According to our study, there is a significant amount of research works on plagiarism detection have been done in English and other languages. Unfortunately, all plagiarism detection tools including renowned Turnitin do not support plagiarism detection in Bangla language where Bangla is the 7th most widely spoken language around the world [3]. Bangla textbooks are freely accessible on different online websites. Contents of these books are regularly getting plagiarized and people behind it getting money and fame instead of the original authors. Thus, we need a fruitful and efficient Bangla plagiarism detection tool to stop Bangla text fraud. This motivates us to develop a Bangla plagiarism detection mechanism.

As we mentioned earlier, there is no work has been done in Bangla plagiarism detection. However, a few related works have been done on Bangla language. In 2017, Nazmul Islam et al. implemented a stylometric approach for detecting authorship from Bangla texts [4]. They classify authorship using n-grams and evaluated their system depends on 3125 passages which are written by 10 Bengali authors. Their system achieved 96% accuracy. In 2012, Momtaz et al. developed an algorithm to detect document similarities for big data using the Ferret model [5]. They used MapReduce and the Hadoop framework. They evaluate their system in standalone and cluster mode. Suprabhat Das and Pabitra Mitra proposed author identification in Bengali literary works in 2011 [6]. They considered only three authors and use unigram and bi-gram features. For large training datasets, the system achieved 90% accuracy for unigram features and 100% accuracy for bigram features. As we mentioned, a significant amount of plagiarism detection tools available in English language and other western languages. We take the help of these papers to improve our understanding of the plagiarism detection system. In 2015, N. Riya Ravi et al. proposed an exploration of Fuzzy C means clustering algorithm in the external plagiarism detection system[7]. As a dataset, they used PAN 2013 corpus for evaluation of their system. The system results are compared with existing approaches, via N-gram and K Means Clustering. They achieved 100% recall values for set 1 and around 98% for sets 2 & 3. In 2011, Martin Potthast et al. explained cross-language plagiarism detection and pointed out a basic retrieval strategy[8]. They survey existing retrieval models in detail and proved the CL-C3G model and the CL-ESA model are better suited for this task. Efstathios Stamatatos developed a monolingual plagiarism detection system which is based on structural information in 2011 [1]. In the same year, Asim et al. have done an overview of comparison between five of the software which are PlagAware, PlagScan, Check for Plagiarism, iThenticate, and PlagiarismDetection.org[9]. In 2009, Markus Muhr et al. proposed a plagiarism detection system based on nearest neighbor search algorithm to treat external plagiarism detection and use the stylometric feature for intrinsic plagiarism detection [10].

In this paper, we have proposed a comprehensive closed domain Bangla plagiarism detection and corpus creation approach. As we have mentioned earlier, to reduce the domain complexity at this stage, we have chosen the educational domain for our corpus. We have developed the corpus with all the textbooks of class I to XII of NCTB except English literature. We tokenize each sentence in both corpus and suspicious text. After tokenization, we have removed the Bangla stop words from each sentence. Then, we generate TF-IDF [11] scores of each sentence in both corpus and suspicious text and finally, we compare these values using Cosine Similarity [12] algorithm to generate similarity. If the similarity is greater than a specific threshold, we consider that portion of text is plagiarized. It also marks plagiarized

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

sentences and indicates the source document. The system shows a complete report of plagiarism of the suspicious document. Since this is the first work on Bangla plagiarism detection, we will publish our project in our GitHub repository including corpus for future researchers. The tool will certainly develop awareness among students and authors.

The paper is organized as follows: In section II, we described our proposed method with step by step explanations. In section III, we described our experimental result and performance analysis of the system. Finally, in section IV, we conclude the paper mentioning the limitations of our system and future work.

II. PROPOSED METHOD

In this section, we have described our methodology to develop the Bangla plagiarism detection system.

A. Corpus Creation and Pre-Vectorization

Corpus is a collection of written texts and most important element of our plagiarism detection tool. We have developed our corpus on the educational domain by downloading all the textbooks of Class I-XII NCTB except English literature books and can be found in [13]. Each of these books requires processing before finally added to our corpus. Preprocessing includes the following steps:

1) Text Extraction: The textbooks are available in Pdf format only. First, we convert all the textbooks into writable Docx format. Then, we remove all the images and tables from the books followed by removal of all additional whitespaces and tabs. This gives us a book with plain Bangla texts only. The whole text extraction process has been done manually and it consumes huge time. Thus, we are developing an autonomous tool to do this task.

2) Tokenization: Tokenization is the process of splitting text based on some delimiter such as punctuation, newline, tab, character etc. Here, we split extracted texts by "।(Bangla full stop)". This gives us separate sentences and we then trim each sentence to remove additional whitespaces. The tokenization part is done automatically by Python's NLTK library functions[14].

3) Stop words Removal: Stop word is a word that does not put any significant information in a sentence[15]. Thus, stop words are filtered out after tokenization and before the processing of natural language data. In Bangla, these are এবং (and), ইহা (it), অনেক (many), তারপর (then) etc. We have created a dataset of Bangla stop words by combining and removing redundant stop words from different sources [16][17][18].

4) TF-IDF Scores Generation: After removing stop words, we have generated the vector representation of the sentences using the TF-IDF algorithm. TF means term-frequency while IDF means inverse document frequency[11]. It is a statistical measure that decides how significant a word is to a corpus or in a single document. Our similarity measure algorithm detects similarity between texts using these TF and IDF scores. Around 83% of the text-based recommendation system uses the TF-IDF algorithm in digital libraries [19]. It uses flowing decision rule,

$$tf - idf(t,d) = tf(t,d) \times idf(t)$$

Here t is a certain term and d is a total number of the document.

The simplified formula of tf(t,d) is,

$$tf(t,d) = \frac{t}{d}$$

Where t is the total number of times a term occurs in a document and d is the total number of the document.

The simplified formula of idf(t) is,

$$idf(t) = log\frac{1+n}{1+df(t)} + 1$$

Where n is the total number of documents, and df(t) is the number of documents with the term in it.

We then store these TF-IDF scores of each sentence with the corpus sentence as Pre-Vectorization technique to make the system faster and efficient than the on the fly vectorization technique. Algorithm 1 demonstrates the procedure of corpus creation in brief.

Algorithm 1 Corpus Creation

```
1: booksPDF ← download all books from NCTB
2: booksPDF← remove all English literature books
3: for each book in booksPDF do
4:     bookDocx ← convert book into writable docx
5:     remove all images and tables from bookDocx
6:     remove whitespaces and tabs from bookDocx
7:     tokenize bookDocx using '।'(full stop)
8:     for each sentence in bookDocx do
9:         trim sentence
10:        remove Stop Words from sentence
11:        TFV ← Calculate TF value from sentence
12:        IDFV← Calculate IDF value from sentence
13:        TFIDF← Calculate TF-IDF score
14:        store TFIDF score in corpus
15:    end for
16: end for
```

B. Similarity Measure:

To detect the plagiarism in a suspicious test. We vectorize the suspicious text in the same way we vectorize reference corpus texts. First, we apply some basic filter on the suspicious text by removing punctuations and additional whitespaces. Then, tokenizing the text by "। (Bangla full stop)" followed by stop words removal and vectorizing the text using TF-IDF algorithm. Finally, we pass this TF-IDF scores of both the suspicious text and the corpus texts to the Cosine Similarity algorithm to generate the similarity score.

Primarily, we have considered two similarity measure algorithms and they are Cosine Similarity and Jaccard Similarity algorithm. After analysis of these algorithms, we found Cosine Similarity gives better performance than Jaccard on Bangla text. We have compared and analyzed the algorithms for the same similarity task. The comparison and analysis are given below.

Article 1: "বাতাসে প্লাস্টিক, খাবারে প্লাস্টিক, এমনকি প্লাস্টিক পানিতেও। আর সেই প্লাস্টিক কণাই প্রতিদিন মানুষের দেহের ভেতরে ঢুকে বিষিয়ে দিচ্ছে শরীর। এতদিন পরিবেশের উপরে প্লাস্টিক দূষণের প্রভাব নিয়ে চিন্তায় ছিলেন বিজ্ঞানীরা।"

Article 2: "খাবারে প্লাস্টিক, বাতাসে প্লাস্টিক, এমনকি প্লাস্টিক পানিতেও। আর এই প্লাস্টিক প্রতিদিন মানুষের দেহের ভেতরে ঢুকে বিষিয়ে দিচ্ছে শরীর। বিজ্ঞানীরা এতদিন পরিবেশের উপরে প্লাস্টিক দূষণের প্রভাব নিয়ে চিন্তায় ছিলেন।"

Table 1: Comparison of performance on Bangla article 1 and 2

Cosine similarity	Jaccard similarity
0.835	0.758

Table 1 shows the similarity measure between Article 1 and 2 by Cosine and Jaccard similarity algorithm. We find out the Cosine similarity gives accurate similarity result than the Jaccard similarity algorithm by manually calculating the approximate similarity between these two articles. Jaccard similarity does not give better outcomes in this situation where the passage is paraphrased by making some structural changes or modifications on words or sentences.

Cosine similarity is a technique of measure similarity between two non-zero vectors. It measures the cosine angle between them using the Euclidean dot product formula where the outcome is bounded between o and 1. Formula is,

$$A \cdot B = |A||B|cos\theta \quad (1)$$

$$similarity = cos\theta = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2)$$

If the similarity score between the suspicious text and corpus text is greater than a threshold, we consider it as plagiarism. In our system, the threshold is 0.7 or 70%. We arrived at this threshold via manual trial and error on a small subset of suspicious documents. When a sentence is flagged as plagiarized, the system stores the sentence, similarity score, source/book name and other necessary information for the demonstration of the final plagiarism report. In the final report, we show the percentage of similarity from each source along with the total similarity of the suspicious document. If the system returns very small similarity from a source or multiple sources it generates a summation of these similarity scores and shows the summation result under the source named "অন্যান্য(other)". Algorithm 2 demonstrates the process of plagiarism detection in brief.

Algorithm 2 Plagiarism Detection

```
1: suspiTxt ← input suspicious text/document
5: remove all images and tables from suspiTxt
6: remove whitespaces and tabs from suspiTxt
7: tokenize suspiTxt using '।'(full stop)
8: for each sentence suspiSen in suspiTxt do
9:     trim suspiSen
10:    remove Stop Words from suspiSen
11:    TFV ← Calculate TF value from suspiSen
12:    IDFV← Calculate IDF value from suspiSen
13:    TFIDF← Calculate TF-IDF score
14:    for each sentence corpSen in corpus do
11:        corpTFIDF ← fetch TF-IDF score of corpSen from corpus
12:        sim ← Cosine Similarity(TFIDF, corpTFIDF)
13:        if sim > threshold  (in our case 0.70) then
14:            flag suspiSen as plagiarized
15:            store suspiSen, sim and other pointers
16:        end if
15:    end for
16: end for
17: display similarity report with all details
```

III. EXPERIMENTAL RESULT AND PERFORMANCE ANALYSIS

In this section, we have demonstrated the experimental result and performance analysis of the system. The experiment has been performed on Kaggle Notebook. Kaggle Notebook is a free cloud computational service with 5 GB disk space, 13 GB RAM and Tesla P100 GPU 16 GB memory [20]. Here, we have shown the performance and output of our plagiarism system with two sample suspicious texts. In suspicious text-1, we have compiled a passage with text from 3 different books. Since the whole passage consists of texts only from textbooks of NCTB, the passage is 100% plagiarized.

Input: *Suspicious text-1:*

"টিকেটের মূল্য অনলাইনেই পরিশোধ করা যায়। এভাবে ইলেকট্রনিক পদ্ধতিতে সেবা প্রদানের ব্যাপারটি ই-সার্ভিস হিসেবে চিহ্নিত করা যায়। একজন আলী সেজে একটি ভাঙা কলসিতে কতকগুলো মাটির চেলাপূর্ণজলপাইয়ের কলসি তৈরি করল। বিচার আরম্ভ হলে আলী নালিশ করল। কাজি বলল,আচ্ছা জলপাই কত দিন পর্যন্ত ভাল থাকে বলো তো?ব্যবসায়ী বলল, যত্রে রাখলে বড়জোর ছয় মাস টাটকা থাকে। ভাবে আর কাজে সম্বন্ধটা খুব নিকট বোধ হইলেও আদতে এ-জিনিস দুইটায় কিন্তু আসমান-জমিন তফাৎ। ভাব জিনিসটা হইতেছে পুষ্পবিহীন সৌরভের মতো, একটা অবাস্তব উচ্ছাস মাত্র। তাই বলিয়া কাজ মানে যে সৌরভবিহীন পুষ্প, ইহা যেন কেহ মনে করিয়া না বসেন। কাজ জিনিসটাই ভাবকে রূপ দেয়, ইহা সম্পূর্ণভাবে বস্তুজগতের। "

Output:

Table 2: Similarity report of suspicious text 1.

Sources	Similarity
আনন্দ পাঠ – ৮ম শ্রেণী	35.80%
সাহিত্য কণিকা – ৮ম শ্রেণী	35.72%
তথ্য ও যোগাযোগ প্রযুক্তি – ৯ম-১০ম শ্রেণী	23.26%
তথ্য ও যোগাযোগ প্রযুক্তি– ৮ম শ্রেণী	3.54%
অন্যান্য	1.68%
Total Similarity	100%

Our system found plagiarism in three major sources as expected for the suspicious text-1 which are " আনন্দ পাঠ – ৮ম শ্রেণী(Anondo Pat Class-8) ", " সাহিত্য কণিকা – ৮ম শ্রেণী (Sahitto Konika Class-8)", and " তথ্য ও যোগাযোগ প্রযুক্তি–৯ম-১০ম শ্রেণী (ICT Book Class-9&10)". Moreover, the system detected one minor source " তথ্য ও যোগাযোগ প্রযুক্তি–৮ম শ্রেণী (ICT Book Class-8)" and some negligible sources combinedly named "অন্যান্য(other)". Although the passage of suspicious text-1 has been created by compiling texts from the first 3 books only, it found a small fraction of similarity from other books as well.

In suspicious text-2, we have compiled a passage where first two lines are from a textbook, next three lines are from another textbook and last ten lines are from a story written by an author of this paper.

Input: *Suspicious text-2:*

"বাঙালির প্রথম যে সাহিত্যকর্মের সন্ধান পাওয়া যায় তা চর্যাপদ নামে পরিচিত। পণ্ডিত হরপ্রসাদ শাস্ত্রী প্রথম নেপালের রাজ দরবার থেকে এগুলো আবিষ্কার করেন। রূপালি স্রোত পাড়ি দিয়ে তারা এক সময় ঢাকায় পৌঁছায়। সেখানে তারা প্রথমেই যায় লালবাগের নল দুর্গে।

এ যেন ফেলে আশা মুঘল সাম্রাজ্যের একটুকরো রাজত্ব। যে রাস্তায় দাঁড়িয়ে ছিলাম তার একটু সামনে গেলেই বা দিকে চিকন একটি রাস্তা চলে গেছে সেই জমিদার বাড়ির দিকে। রাস্তাটা ঘাস আর বন্য লতা-পাতায় ভরা, অনেক দিন যে কেউ এই পথ মাড়ায়নি তা ভালোই বুজা যাচ্ছে। দুপাশে ধানখেত, একটু হাটার পর ওই জমিদার বাড়ির কাছে চলে আসলাম। এদিকে খুব ঘন জঙ্গল,বড় বড় গাছে ভরা। চাদের আলোর লেশমাত্র নেই এদিকে। স্তব্ধ,অন্ধকার, প্রাণহীন এক জঙ্গল। হারিকেনের আলোটা বাড়িয়ে দিলাম। হঠাৎ বনের ভিতর থেকে একটি পাখি বিকট উদ্ভুত একটা শব্দ করে ডেকে উঠলো। তিথি খপ করে আমার বা হাতটা চেপে ধরল। মনে মনে আতকে উঠলেও, তিথিকে বুজতে দিলাম না।"

Output:

Table 3: Similarity report for suspicious text-2.

Sources	Similarity
চারুপাঠ- ৬ষ্ঠ শ্রেণী	16.90%
বাংলাদেশ ও বিশ্বপরিচয়-৮ম শ্রেণী	15.49%
অন্যান্য	0.8%
Total Similarity	33.19%

Table 3 shows two major sources of similarity which are " চারুপাঠ- ৬ষ্ঠ শ্রেণী (Charupath Class-6)" and " বাংলাদেশ ও বিশ্বপরিচয়-৮ম শ্রেণী (Bangladesh o Biswa Parichay Class-8)". This shows the passage has a similarity score of 33.19% which is correct since the majority of the portion is genuine texts written by an author of this paper as we have mentioned earlier.

We have manually tested our system with a test dataset of 400 Bangla passages consists of fully copied passages, partially copied passages and not at all copied or genuine passages. The test result shows, in 386 passages it gave 100% accurate report and in 14 passages it provided partially incorrect report. This gives us 96.75% overall accuracy by our Bangla plagiarism detection tool.

IV. Conclusion and Future Work

In the age of information technologies, plagiarism has become a threat and turned into a serious problem to the genuine authors. The paper demonstrates every step of our closed domain Bangla plagiarism detection and dataset creation approach. We have created a dataset which consists only textbooks of NCTB at the current state. We have successfully implemented the plagiarism detection tool for Bangla language and tested with several Bangla passages. It shows a significantly high accuracy of 96.75% at the current approach and corpus.

Although we have achieved our goal, the system has some limitations as well. According to our study, the first major limitation is the size of our corpus and the second is handling the paraphrased sentences. At current state, the system is based on a dataset which consists only textbooks. Moreover, as we have mentioned earlier, we use TF-IDF for vector representation from a string. Here TF-IDF has some limitations such as it works based on term and number of existences of a particular word which is weak handling synonyms. This leads us in a difficult situation to detect paraphrased sentences. Authors are planning to use Deep Neural Networks approach to improve the performance of our plagiarism detection system. Moreover, authors are working to make the system more robust so that it can detect Bangla difficult paraphrased sentences as well. Besides, authors will increase the dataset to cover more domains shortly.

V. References

[1] E. Stamatatos, "Plagiarism detection based on structural information," in Proceedings of the 20th ACM international conference on Information and knowledge management, 2011.

[2] S. M. a. S. B. Zu Eissen, "Intrinsic plagiarism detection," in European conference on information retrieval, 2006.

[3] Ethnologue, "Ethnologue Languages of the World," 2019. [Online]. Available: https://www.ethnologue.com/guides/ethnologue200.

[4] N. a. H. M. M. a. H. M. R. Islam, "Automatic authorship detection from Bengali text using stylometric approach.," in 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2017.

[5] A. a. A. S. Momtaz, "Detecting document similarity in large document collecting using MapReduce and the Hadoop framework," 2012.

[6] S. D. a. P. Mitra, "Author identification in Bengali literary works," in International Conference on Pattern Recognition and Machine Intelligence, 2011.

[7] K. V. D. G. N. Riya Ravi, "Exploration of fuzzy C means clustering algorithm in external plagiarism detection system," in Intelligent Systems Technologies and Applications, 2016.

[8] M. a. B.-C. A. a. S. B. a. R. P. Potthast, "Cross-language plagiarism detection," Language Resources and Evaluation, vol. 45, no. 1, pp. 45--62, 2011.

[9] A. M. E. T. a. A. H. M. D. a. S. V. Ali, "Overview and Comparison of Plagiarism Detection Tools," DATESO, pp. 161--172, 2011.

[10] M. Z. R. K. a. M. G. Markus Muhr, "External and intrinsic plagiarism detection using vector space models," in Proc. SEPLN, 2009.

[11] Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

[12] Rahutomo, Faisal & Kitasuka, Teruaki & Aritsugi, Masayoshi. "Semantic Cosine Similarity". The 7th International Student Conference on Advanced Science and Technology ICAST (2012).

[13] http://www.nctb.gov.bd/site/page/06666571-1632-4dd4-b534-0a7450947e3b/-

[14] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.

[15] A. Rajaraman. J.D. Ullman, "Data Mining". Mining of Massive Datasets, Cambridge, England, CUP, 2011, ch 1, pp. 1-17.

[16] https://www.ranks.nl/stopwords/bengali

[17] https://github.com/stopwords-iso/stopwords-bn

[18] https://sanjir.com/6202/

[19] S. L. M. G. B. G. C. B. a. A. N. Joeran Beel, "Research paper recommender system evaluation: a quantitative literature survey," in Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, New York, 2013

[20] "Notebooks Documentation." Kaggle, www.kaggle.com/docs/kernels ?fbclid=IwAR0a4pxbvpQetZ6NO8rvsLXQ9SOzcqCJG1a4DP8N-Dy-ZzOfxrFDwdYkwrM#technical-specifications.