Identify Plagiarism in Bangladeshi News Article with Corpus Creation Approach

Jonayed Hossen Antor

A Thesis in the Partial Fulfillment of the Requirements

for the Award of Bachelor of Computer Science and Engineering (BCSE)

Department of Computer Science and Engineering
College of Engineering and Technology
IUBAT – International University of Business Agriculture and Technology

Fall 2019

Identify Plagiarism in Bangladeshi News Article with Corpus Creation Approach

Jonayed Hossen Antor

A Thesis in the Partial Fulfillment of the Requirements for the Award of Bachelor of Computer Science and Engineering (BCSE)

The thesis has been examined and approved,

Prof. Dr. Utpal Kanti Das
Chair and Professor

Dr. Hasibur Rashid Chayon
Co-supervisor, Coordinator and Associate Professor

Md. Sakibul Islam
Supervisor and Lecturer

Department of Computer Science and Engineering
College of Engineering and Technology
IUBAT – International University of Business Agriculture and Technology

Fall 2019

Letter of Transmittal

30 June 2021
The Chair
Thesis Defense Committee
Department of Computer Science and Engineering
IUBAT–International University of Business Agriculture and Technology
4 Embankment Drive Road, Sector 10, Uttara Model Town
Dhaka 1230, Bangladesh

Subject: Letter of Transmittal.

Dear Sir,

With due respect we the students of BCSE are going to hand over you our thesis report based on Plagiarism detection in newspaper articles.

Though we are in our learning curve, the report has enabled us to learn new technology by this work. We have given our best throughout the work. Thank you for your supportive consideration about our work. Without your inspiration this report would be an incomplete one. Thank you for your support and cooperation.

Yours sincerely,

Jonayed Hossen Antor
ID-16203062

Student's Declaration

We declare that this thesis is an original report of our research, has been written by us after doing the registration for the degree of BSc in CSE at International University of Business Agriculture and Technology, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or any qualifications. The experimental work is almost entirely our own work; the collaborative contributions have been indicated clearly and acknowledged. Due references have been provided on all supporting literatures and resources.

Also, we have read the University's research guidelines, and accept responsibility for the conduct of the procedures in accordance with the University's thesis Committee. We have attempted to identify all the things related to this research that may arise in conducting this research, obtained the relevant ethical approval, and acknowledged our obligations and the rights of the participants in this research.

Jonayed Hossen Antor
ID-16203062

Supervisor’s Certification

This is to clarify that the student Jonayed Hossen Antor (ID #16203062) carried out their thesis work “Identify Plagiarism in Bangladeshi News Article with Corpus Creation Approach” at International University of Business Agriculture and Technology (IUBAT) Fall 2019. During this period, they consulted me on regular basis as required by the department. I therefore recommend that theirthesis report be prepared for final presentation.

Md. Sakibul Islam

____

Supervisor and Lecturer

Department of Computer Science and Engineering

IUBAT–International University of Business Agriculture and Technology

Abstract

The word plagiarism has come from the Latin word 'plagiarius,' that means to kidnap. Plagiarism means stealing and passing off ideas of another. Plagiarism is presenting someone else’s work or ideas as your own, with or without their permission, by integrating it into your work without full credit. Plagiarism includes copying the existing information in adapted format or sometimes the original article as it is. Plagiarism is a corruption. Plagiarism can be a violation of copyright laws and can be considered deceitful. All published and unpublished material, whether in document, printed or electronic form, is protected under this definition. Plagiarism may be planned or uncontrolled, or accidental. Under the guidelines for examinations, planned or uncontrolled plagiarism is a disciplinary crime. There are numerous tools available for identifying plagiarism in the English language and other western languages. There is no available news article plagiarism identify approach in Bangla language where Bangla is one of the most spoken languages in the world and widely used on the internet. Therefore, we have decided to develop a plagiarism identification tool for Bangladeshi news articles with corpus creation approach.

Acknowledgments

First, we have to thank our thesis supervisor, [Md. Sakibul Islam]. Without his assistance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. We would like to thank you very much for your support and understanding in our research work.

We would also like to acknowledge our batchmates, seniors and all those have given us support and encouraged us in our research work. We are really thankful to them.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Background and Context

Plagiarism is stealing and passing off or we can say that it is a way to represent others writing, any work, thought, ideas, languages as their own work. It is a form of intellectual stealing and scam. It is a violation of copyright laws. It presents as new and original thought derived from an existing one. Plagiarism can be deliberate or reckless, or accidental. In simple word we can say plagiarism is an act of fraud.

There are different types of plagiarisms. For examples:

- Other's work can be turned into your work without knowing it
- Copy huge parts of text from a source without mentioning that source
- Copy passages from different sources, patching them together, and turning in the work as your own.
- Copy from a source but change a few words and phrases to disguise plagiarism.
- Paraphrase from a number of different sources without citing those sources.
- Turn in work that you did for another class without getting your professor's permission first.
- Buy an essay or paper and turn it in as your own work.
- Mention a writer or source within your paper without including a full citation in your bibliography.
- Quote a source with incorrect information, making it difficult to find that source.

- Use a direct quote from a source, quoting that source, but failing to put quotation marks around the copied text.
- Paraphrase from multiple cited sources without including any original work.

In Plagiarism, there are different field in it. It is common scenario to find plagiarism stories about students, writers, historians, and songwriters that they are plagiarizing the work of others. But, most alarmingly for journalists, there have been a number of high-profile cases in current years of plagiarism by reporters.

If one reporter or writer writes an article in a distinct way or a specific way and another reporter copy passages from earlier story, then it will be called plagiarism in news article writing. A writer decides that he wants to make an online website to generate ad revenue. Instead of writing his own articles, he visits other websites that have articles on the topic in which he is interested. He copies each of the articles, changes the titles and the writer's names to his name and posts the articles on his own website.

Plagiarism can take different forms in journalism. For Examples:

- Data: This involves using information that another journalist has gathered without giving credit that information to the journalist or to his or her publication.
- Writing: If a journalist writes a story in a unique or unusual way, and another journalist copy passages from that story into his own article, this is also an example of plagiarism.
- Ideas: When a journalist, usually a columnist or news analyst, advances a novel idea or theory about an issue in the news, and another reporter copies that idea then this idea plagiarism happened.

Using plagiarism detection tool to identify plagiarism in article is the best way to avoid plagiarism in article writings. Plagiarism detection tool can give proper result within a second by checking plagiarism in whole article. It helps to write a unique article by establishing the authenticity of work. All the checker tool their own unique algorithms to find out plagiarisms from large chunks of texts against indexed web pages.

1.2 Scope of Our Topic

Many people act like plagiarist in the field of News Media in Bangladesh. They search relevant article from a news portal, then copy and publish that news without giving any credit to the main authors. There are numerous methods available for detecting plagiarism in the English language and other western language. However, there is no plagiarism detection tool for identify any plagiarism in our Bangla news articles. So, we thought to develop a plagiarism detection tool with corpus creation for Bangla News Articles.

1.3 Evaluation of Current Situation

According to our study, there is a significant amount of research works on plagiarism detection have been done in English and other languages. But 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.

Bangla Newspaper Articles are freely accessible on different online News Portal. Contents of these articles are regularly getting being plagiarized and people behind it getting money and fame instead of the original writers. Thus, we need an efficient Bangla plagiarism detection tool to stop Bangla text fraud. This motivates us to develop a Bangla plagiarism detection tool.

1.4 Importance of Our Work

- As there is no available news article plagiarism identify approach in Bangla language; our work will help the news agency as well as the writers.
- In our work, links of source article text will be provided to the users.
- It will be useful in writing original work within a short time.
- It will help in identifying who copies whom.
- It will also ensure the paper quality and originality of the work.
- Thus, it may be helpful to prevent plagiarism from our news media.

1.5 Research Problem

- There is no available news article plagiarism identify approach in Bangla language.
- Writers found problems in writing original work within a short time.
- There is no available tool for identifying who copies whom.
- People cannot identify the original work.
- Original writers don’t get fame and name by their writings, on contrary plagiarist is getting fame and money both.

1.6 Research Aim

- It will ensure that writer’s own news article is free from plagiarism.
- It will reveal plagiarism between two news articles covering same event.
- It will be free in use. There will be guarantee of accuracy and security of news articles. Users will be able to find a huge number of news article at once.

- The main writer of the article he will be benefited. Because it will ensure the originality of the work. And whoever wants to write something will also know that their article is plagiarism free or not.

In this paper we proposed an extrinsic plagiarism detection in news articles. We have divided our work in two parts. Firstly, we are going to build a corpus as our dataset. As we are working with news article, when there will be lots of images, tabs and others things. We will store only text. So, we have to remove those extra things. So, we have pre-process data. Data pre-processing can be done by 5 things. Text Extraction, Tokenization, removing stop words, removing punctuations and removing null values. And then we are going to develop plagiarism detection method dedicated to Bangla language. We are going to use TF-IDF algorithm with Bag of Words and N-Grams we will generate pre vectorization. But TF-IDF contains information on the more important words and the less important ones as well. So, we will use Bag of words method with TF-IDF. Then using Cosine similarity algorithm, we will generate similarity.

The rest of the paper has been organized in some sections. The section 2 provides a background and discusses related literature. The section 3 introduces the research methodology. The section 4 discusses the experiments and findings. Finally, the section 5 presents the conclusion, system limitations and future works of this paper.

2. Literature Review

Machine learning is the study of computer algorithms that improve automatically through experience. It is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. Plagiarism Detection is a non-trivial task and is still an active area of research. Machine learning can predict plagiarism with a reasonably high degree of accuracy. Many machine learning algorithms are applied in plagiarism detection. These are:

- TF-IDF Algorithm
- Cosine similarity Algorithm
- Jaccard similarity Algorithm
- Hashing Algorithm
- Clustering Algorithm
- Spatio Algorithm
- Temporal Algorithm
- SCAM Algorithm
- Nearest neighbor search algorithm

As we mentioned earlier, there is no work has been done plagiarism in Bangla news articles. However, a few related works have been done on Bangla language.

I. In 2020, Adil Ahnaf et al. developed first ever plagiarism tool for Bangla language. They constructed the corpus with all the textbooks of class I to XII of NCTB except English literature. They used closed domain for creating dataset. And included only the NCTB textbooks for reducing the domain complexity. They tokenized and removed the Bangla stop words from each sentence in both corpus and suspicious text. Then, they generate TF-IDF scores in both corpus and suspicious text and finally, they compare these values using Cosine Similarity algorithm to generate similarity. If the similarity is greater than a specific threshold, they consider that portion of text is plagiarized. It also marks plagiarized sentences and indicates the source document. The system shows a complete report of plagiarism of the suspicious document.

They have used:

- Closed domain, for building a dataset, corpus creation
- TF-IDF Algorithm
- Cosine similarity and Jaccard similarity Algorithm
- Stylometric approach
- Kaggle Notebook (for free cloud computational service)

They have added in their corpus only the NCTB books to reduce their domain complexity. But by reducing their domain complexity they have made a small corpus. It cannot handle the paraphrased sentences and also weak in handling synonyms. It is limited for the users.

II. In 2009, Hwan-Gue et al. developed plagiarism detection algorithm for identifying plagiarism between news articles covering one event. They used Spatio-Temporal evolution Model to identify plagiarism. Here they constructed an evolution tree, there they can find documents for detecting plagiarisms. In document evolution tree, nodes denote documents and edges show relationships between two documents. They used ERP values for evolution tree and kruskal's MST algorithm for constructing tree.

They used:

- Spatio-Temporal evolution Model
- Biological aspects for evolution analysis
- In-Silico evolution process
- Evolution plagiarism probability
- DEVAC system
- Kruskal's MST algorithm for constructing tree.

They have introduced spatio temporal model together but this model isn't able to find who copies from whom. However, it is quite complex structure to understand and it is not an automated system.

III. In 2006, Wolfgang et al. used prototype system and diffusion of textual content. This paper is based on Austrian Press Agency (APA).This is a semi-automatic identification. There is a generic database technology of APA which is PowerSearch. Server functionality is available for retrieval of documents based on Boolean quarries and results show in topical clusters. Dataset showed the results of last 30 days. A clustering Algorithm is used provided by PowerSearch database; based on scatter-gather approach. But this algorithm computed similarities relate to an article as a whole. So they find a more suitable syntactical clustering which relies on hashing of document sketches. They have used hashing algorithm (simple MD5 hash). But there is no dataset for define statements.

They have used prototype system and diffusion of textual content in plagiarism detection with a semi-automated system but there is no sample dataset for slogans or defined statements.

IV. In 2019, Santipong et al. developed this corpus by using simulated plagiarism method based on Thai Wikipedia articles and web pages articles. They have used Thai plagiarism annotation tool, Assist human annotators. It is based on four classes of Thai plagiarism and linguistic mechanism. These are: Copy based change, Lexicon based change, Structure based change, and Semantic based change. Corpus is manually created by human consisting of 100 documents of length between 200-300 words. It is created with four levels of reuse and those are: Near copy, Light revision, Heavy revision, and on-plagiarism. It has 5 source documents. It has adopted three main classes of framework.

They have used:

- Thai plagiarism annotation tool
- Assist human annotators
- Consist of three steps
- Selecting relevant source documents
- Copying texts
- Obfuscating text

They have developed an effective algorithm for identify plagiarism but corpus size small. It is not supported by any other languages. There is no automatic generation in plagiarism cases.

V. In 2017, Mahmoud and Zrigui proposed a system for detecting semantic plagiarism in Arabic documents that benefited from machine learning technology. In the preprocessing phase, the suspicious and source documents were split into sentences then into words without removing stop words. In the feature extraction phase, the TF*IDF measure was calculated for weighting words in terms of importance. Then the word2vec algorithm was used for learning word embedding, and the skip-gram model was employed for predicting the context of words given a current word vector. They used cosine and the Euclidean distance measures for similarity calculation. The degrees of similarity between sentences were compared to a predefined threshold. Experiments were conducted on an open-source Arabic corpus and they claimed a precision rate of (85%) and a recall rate of (84%).

They used a Convolutional Neural Network (CNN) approach for detecting paraphrasing plagiarism in Arabic documents. This method is supposed to detect paraphrasing plagiarism through the measurement of semantic relatedness between the suspicious and the original documents. This Project has three phases approach: preprocessing, feature extraction, and paraphrase detection. The feature extraction phase employed a skip-gram model for word-to-vector representation after preprocessing, where each document is represented by a vector in a multidimensional space. The paraphrase detection phase applied the cosine similarity measure on the vectors of both the suspicious and the original documents to reduce dimensionality. In the last, one mathematical function called SoftMax was used for paraphrase detection according to some predefined threshold. These experiments showed a precision rate of (88%).

However, they conducted their experimentation on an open-source Arabic corpus, named OSAC (Saad & Ashour, 2010). The corpus was prepared in ten different categories collected from multiple websites. These sources of the articles were news channels and social and commercial websites, which clearly makes it inappropriate for academic plagiarism detection. Specialized content is what the PD corpus must have to consist of, because academics do not normally plagiarize the news or social media.

VI. In 2018, Khorsi et al. developed a Two-Level Plagiarism Detection System, which is supposed to detect different plagiarism cases, which includes verbatim and paraphrasing. This system consisted of two consecutive modules; these are: fingerprinting and word embedding detection. The first module is accountable for preprocessing and segmenting the suspicious document into sentences. When sentences exceeded some threshold value, they were passed on to the second module to test for paraphrasing and synonym replacement. The fingerprinting was applied by chunking the text documents into n-grams and then selecting the least frequent ones. Finally, authors used a function called Brian Kernighan and Dennis Ritchie (BKDR) for hashing the selected n-grams. The first module applied Jaccard measuring similarity, at the same time as the second module used the cosine similarity measure. Important words were picked on the basis of their IDF value and their part of speech tags. To test their approach, Khorsi et al. (2018) used the ExAraDet-2015 corpus. Experimental results showed an overall precision rate of (85%) and a recall rate of (87%).

VII. In 2017, Shanta Paul et al. developed a new corpus of 3000 passages written by three Bengali authors, which was an end-to-end system for authorship classification based on character n-grams. They have removed poetry and songs then merged the remaining prose in a single large file and sampled 25 random Fragments. They have tried three feature representations on the above three categories – binary, TF, TF-IDF. They have found accuracy on held out data reaches 98%; which was obtained for 300-character bigrams (TF) feature combination and naive bayes classifier.

3. Research Methodology

3.1 Research Design

We are developing plagiarism detection system in news article with corpus creation approach. To develop the system at first, we will develop a corpus dataset of the news articles where we can collect & store all the data. When an article will be inputted in the system it will be stored in this corpus. That is called corpus dataset. Before adding data in corpus dataset, the system will pre-process it. We need to pre-process the data because, there have some unnecessary things like; photos, whitespace, stop words. This will make the data large and will take more space. This is why the system will go through this process. To pre-process the data, we have to do Text Extraction, Tokenization, Removing stop words. After doing these things, using tf-idf algorithm with bag of words and n-gram the system will generate vector representation. It is a statistical measure that decides how significant a word is to a corpus. And our similarity measures algorithm will detect similarity between texts using these TF and IDF scores. After doing this the tf-idf the system will compare these values using Cosine similarity algorithm to generate similarity to measure that, the text is similar to the other text or not and it is plagiarism free or not, and will show the percentage of similarity. We will sort the similarity in descending order to show the data with maximum similarity at first.

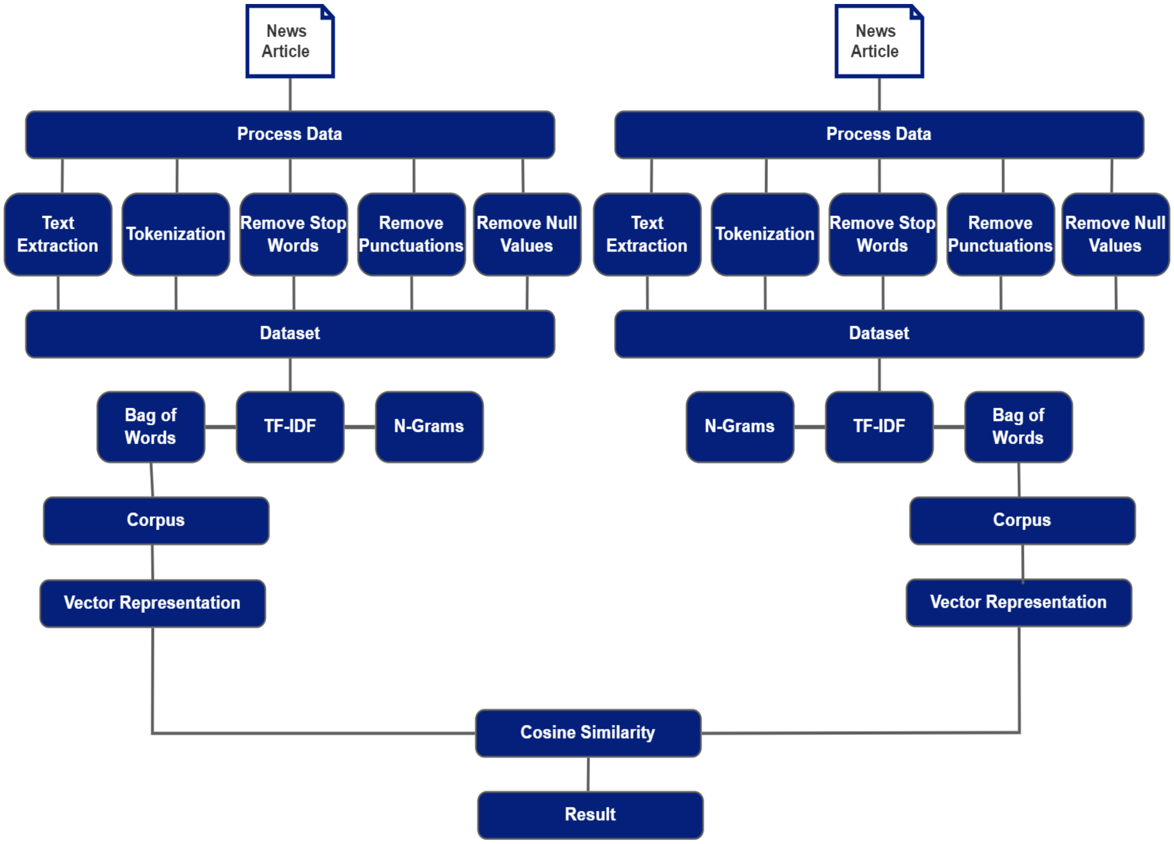


Figure 3.1: System Model for Identifying Plagiarism in Bangladeshi News Article with Corpus Creation Approach.

3.2 Methods and Sources

In our system we are going to use several methods, algorithm and sources to do develop the system. These are:

❑ Corpus Dataset

❑ Data Pre-Processing

- Text Extraction
- Removing Null Values
- Removing Punctuations

    - Removing Stop Words
- ❑ TF-IDF Algorithm with
    - Bag of Words
    - N-Gram
    - Tokenizer
- ❑ Cosine Similarity

Corpus is a dataset where we can store our data and we can use the data as well when needed. As we are working with news article, where there will be lots of images, tabs and others things. We will store only text. So, we have to remove those extra things. Because, those extra things will take more storage and it will make the process slow. This is why we need to remove those things. So, to do this, we have to pre-process data. Data pre-processing can be done by 5 things. Text Extraction, Tokenization, Removing Stop Words, Removing Punctuations and Removing Null Values. Text Extraction is a process where we will remove all the images, tables, additional white space, tabs to make the size smaller. By text Extraction we can split text based on some delimiter such as punctuation, newline, tab, character etc. But as we are working with Bangla news article, so we will split text according to the indexing. Then we will check each text line by line according to their index and will keep this into a variable. If there are any null values, we will drop it. Then, we will check each text character by character and if there is any punctuation or not. If we get any punctuation then, we will remove that and take rest of the text and will keep it into a variable. Then we will remove Stop Words. Stop words are commonly used words and it does not put any significant information in a sentence. So, this not important for us and we will remove it from text. So, it will decrease the corpus size. And it will be easy and less time consuming for preprocessing. And we can save more space.

Then using TF-IDF algorithm with Bag of Words, N-Grams and Analyzer we will generate pre vectorization. TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. But TF-IDF contains information on the more important words and the less important ones as well. Because of this it makes the process slower. To solve the problem, we will use Bag of words method with TF-IDF. This method creates vocabulary of all unique words. And it makes the process faster to detect plagiarism. And N-Gram is a process where we can move text from a sentence.

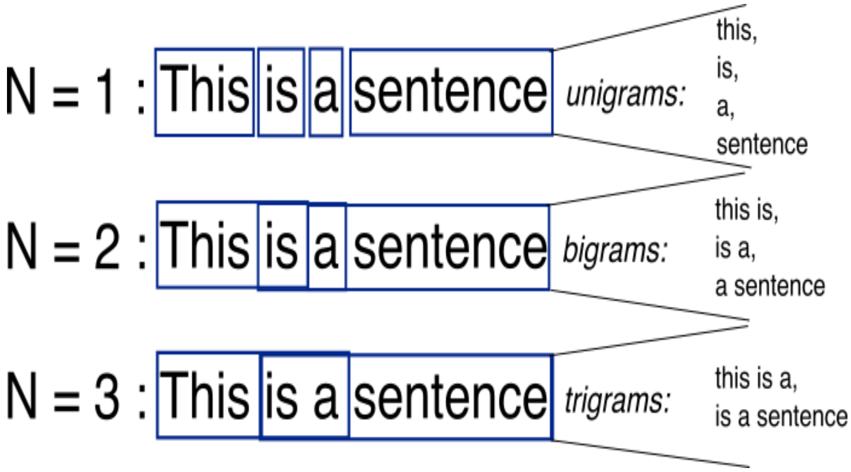


Figure 3.2: N-Gram

On the other hand, we can say, N-Grams are simply all combinations of adjacent words or letters of length n that you can find in your source text. So, inside of TfidfVectorizer we have taken 1 to 3 grams of N-Grams with maximum 5000 features and have taken “word” as analyzer. Then created an instance of TfidfVectorizer and made a matrix with it.

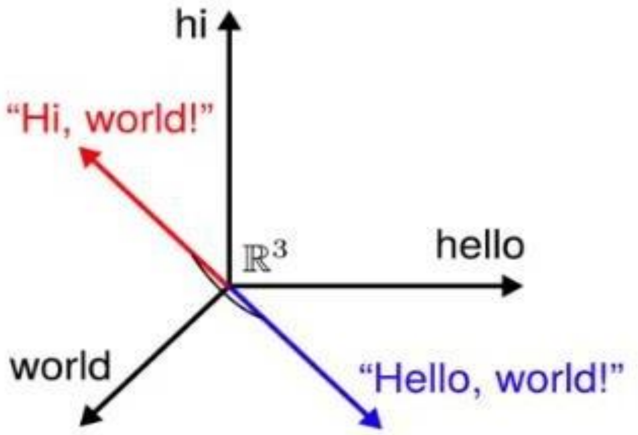


Figure 3.3: Cosine Similarity

Then using Cosine similarity algorithm, we will generate similarity. Cosine similarity algorithm is basically used to generate similarity between two articles. With Linear Kernel using the matrix calculation we are checking the cosine similarity between articles and showing the percentage of the similarity to know it is plagiarized or not. Or, how much the text is similar to the existing data. And data with maximum similarity will comes first. So, these are the methods, algorithm and sources we are going to use.

3.3 Practical Considerations

- As our system is based on news article there have lots of data. It is not possible to store all the things. So, we can store data of limited time that recently published. Because, in news no one copy old articles, they copy recent article. So, by storing recent article we can reduce corpus size.
- For processing huge amount of data needs more RAM space. RAM space of our PC is not stable. So, we have done our project in Google Colaboratory. Google Colaboratory giving 12 GB of RAM space for free to use from cloud.

- We have chosen TF-IDF algorithm because it is easy to use for beginners than other algorithms.
- To detect similarity between articles, there have two types of Algorithm. Cosine Similarity and Jaccard Similarity. As, Cosine similarity shows better result so, we have decided to use Cosine Similarity.

4. Result and Discussion

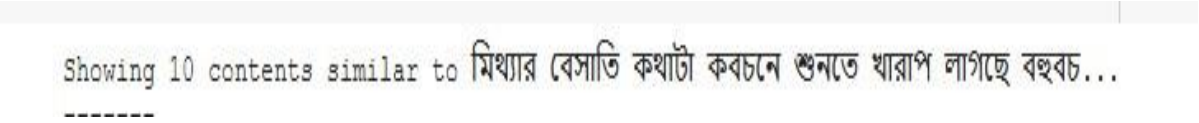
To get the result of plagiarism we have to go through some process like; Tokenization, Removing Stop Words, Removing Punctuations and Removing Null Values, TF-IDF Algorithm, Bag of Words, N-Gram, Cosine Similarity. After doing these processes we can get our result and we can measure the similarity and get the percentage that, this news is plagiarism free or not.

After doing the cosine similarity to show the result of the similarity we have created unique id for each data. Because, we will check the result according to their unique id. Then, inside result function we have sort the cosine similarities in descending order. The reason is, normally the percentage will be shown from minimum to maximum. But we need to see the maximum plagiarized data which is much similar to the inputted data. So, this is why we are showing the result in descending order.

For printing the result, we are printing the inputted data and plagiarized data both. First, we are printing the inputted data. Secondly, we are printing the plagiarized data according to their id from the corpus dataset, that are similar to that inputted data. Also, we are showing the percentage of the plagiarism that, these data are that much similar to that data. We can also set the number of printable data to check the top results.

In this system, we are going demonstrate the experimental result and performance analysis of the system. The experiment has been performed on Google Colaboratory. The system will show the performance and output of our plagiarism system with three sample suspicious texts. In suspicious text-1, we will input a data from its own existing dataset and we will check the similarity of that data. We can input data according to the item id of that data. So, I am inputting the data of item id: 1.

Input:



```
Showing 10 contents similar to মিথ্যার বেসাতি কথাটা কবচনে শুনতে খারাপ লাগছে বহুবচ...
-------
```

Figure 4.1: Inputted data of suspicious text-1

Output:

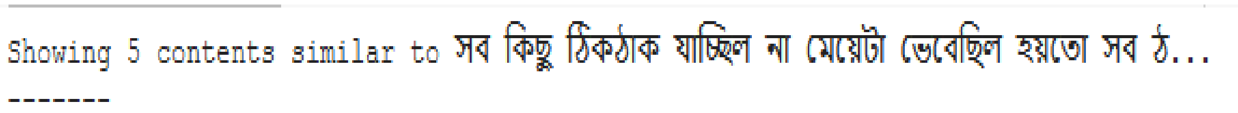
Table 4.1 Similarity report of suspicious text-1

Top Plagiarized Data	Similarity
অরিন খান আইনজীবী শিক্ষাবিদ মানবাধিকার কর্মী রোমভিত্তি...	(score:32.5) %
সমস্যা চিহ্নিত বের বাংলাদেশের মানুষ সত্যি রাজনীতিত...	(score:31.81) %
অধ্যাপক ডি পিটার আগিন দশক অর্থনৈতিক উন্নয়ন বৈশ্বিক...	(score:27.31) %
প্লিমথে দুবার গেছেন মিল্টন তৃতীয়বার গিয়ে মুগ্ধতা...	(score:27.01) %
সৃষ্টির সূচনা প্রাণীর প্রধানতম তাগিদ নিজেকে বাচিয়ে...	(score:26.92) %
ক অর্থনীতি অযৌক্তিক বাদানুবাদে সোজা বাংলায় ড়ে তর...	(score:26.74) %
বনে জঙ্গলে বাস আলাদা কথা সংসারে বাস প্রতিপক্ষের সম...	(score:26.13) %
দেশের সত্যিকার অর্থে উন্নয়ন অবশ্য গ্রামীন পর্যায়...	(score:25.67) %
বাংলাদেশের রাজনীতি ক দিক নির্দেশনাহীন পথে যাত্রা ...	(score:24.68) %
৩৩ টি দেশের জরিপ চালিয়ে রাজনীতিতে অংশগ্রহণের বাংলাদে...	(score:24.61) %

The system is printing the inputted data "মিথ্যার বেসাতি কথাটা কবচনে শুনতে খারাপ লাগছে বহুবচ...." as input. Here, we are printing first 50 characters of that inputted data. Then, we are getting output of top plagiarized data. We are showing top 10 results of plagiarized data that is similar to the text. Also, for outputted data we are showing first 50 characters of each data with the percentage.

In the first one, we showing the plagiarism between its own dataset's data. But if we input a random data to the system to check if the data is plagiarized or not. And showing the possible plagiarized data. We can also do this by same process. Like previous one, first doing preprocessing, then doing the cosine similarity and printing the result in descending order. So, in suspicious text-2, we will input a random data and check the similarity from existing dataset.

Input:



```
Showing 5 contents similar to সব কিছু ঠিকঠাক যাচ্ছিল না মেয়েটা ভেবেছিল হয়তো সব ঠ...
-------
```

Figure 4.2: Inputted data of suspicious text-2

Output:

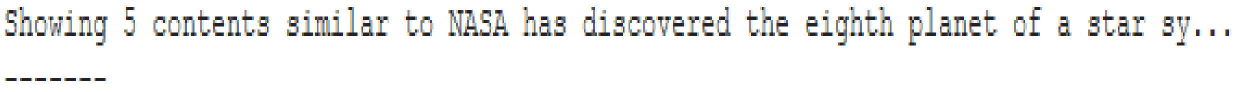
Table 4.2 Similarity report of suspicious text-2

Top Plagiarized Data	Similarity
হাতে সময় সাত দিন পয়লা বৈশাখ নববর্ষ উপলক্ষে বাড়ির সামনে...	(score:27.73) %
জীবনসঙীর কর্মজীবী নারীর অভিযোগ ...	(score:26.93) %
খেলাঘরে বন্ধুরা প্রিয় সংঘটনী...	(score:24.19) %
গল্প বলব বানানি অন্যের মুখে শোনা অপরের মাল নামে চাল...	(score:22.85) %
গল্পটা আগের গল্পটা নিয়ে বড় হয়েছি গল্পটা দর্শকদের আ...	(score:22.25) %

The system is printing the inputted data “সব কিছু ঠিকঠাক যাচ্ছিল না মেয়েটা ভেবেছিল হয়তো সব ঠি...” as input. Here, we are printing first 50 characters of that inputted data. Then, we are getting output of top plagiarized data. We are showing top 5 results of plagiarized data that is similar to the text. Also, for outputted data we are showing first 50 characters of each data with the percentage.

In this system, if we give input of a random English data, we won’t get any result. And, percentage of similarity will be 0.0%. Because, our corpus dataset is in Bangla Language. This is why, if we give any English data, it will not generate any plagiarism. So, in suspicious text-3, we will input a random English data and check the similarity from the existing dataset.

Input:



```
Showing 5 contents similar to NASA has discovered the eighth planet of a star sy...
-------
```

Figure 4.3: Inputted data of suspicious text-3

Output:

Table 4.3 Similarity report of suspicious text-3

Top Plagiarized Data	Similarity
আইসলেন্ডে বছরের সময়ে দীর্ঘ দীর্ঘ পবিত্র রমজানের স...	(score: 0.0) %
ড আনুন নিশাত বর্তমানে ব্রাক বিশ্ববিদ্যালয়ের সেন...	(score: 0.0) %
জিয়া অরফানেজ ট্রাস্ট দুর্নীতি মমলায় বিএনপি চেয়ার...	(score: 0.0) %
আওমীলীগের সাধারণ সম্পয়াদক ওনায়দুল কাদের বলে...	(score: 0.0) %
সৃজনশীল পদ্ধতীতে সমাজকর্ম পরী...	(score: 0.0) %

The system is printing the inputted data “NASA has discovered the eighth planet of a star sy...” as input. Here, we are printing first 50 characters of that inputted data. But in output the percentage of similarity is 0.0%. Because, our dataset is in Bangla Language. So, we give any Other Language data except Bangla Language, we won’t get any result. This is how our system works.

5. Conclusion

Day by day plagiarism is becoming a threat for our news articles writer. Because, most of the time, the real writing is being copied and the others people are getting benefited instead of the real one. We are going to construct plagiarism detection approach for Bangla newspaper articles. This plagiarism identification system will help our news agency to overcome this problem to make the real writing beneficial. And this project will show every step of making difference between genuine writer and plagiarist. Our system will help to find the plagiarist easily for the news agencies and original writers. They easily will find out who copies from whom. It will useful in writing the original work fast and it will ensure the quality of writer's work. It will show the percentage of plagiarism. We are going to test Bangla newspaper articles in our system.

Although we are going to achieve our goal, our project will have some limitations in its system. Storing news article means a lot of data. As much we add more data the dataset will become slow and it will take a lot of storage. So, it will take more time to analyze as much as data we store. And, to process huge of data we need more RAM space. Currently, we have RAM limitation as well. This is why are going to work with using limited data. If we can build a generic dataset instead of corpus dataset it will make the process faster and we can store more data. So, we are planning to make it as our future work. At present we are going to add some preset data in our corpus. As, it takes much time to collect a lot of news article, we have collected a 40k newspaper article dataset from Kaggle. Also, we are using TF-IDF algorithm in our system. 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. By using this we cannot also identify paraphrased sentences as well. We are planning to add word2vec algorithm as well as convolutional neural network (CNN) in our system. And we will also try to add a generic dataset to our system. And we will add more dataset to our corpus and more RAM space to the system to make the system more efficient, faster, accurate and more user friendly to the users.

References

Adil Ahnaf, Nahid Hossain, Shadhin Saha. (5-7 June 2020), Dhaka, Bangladesh. “Closed Domain Bangla Extrinsic Monolingual Plagiarism Detection and Corpus Creation Approach”, in 2020 IEEE Region 10 Symposium (TENSYMP).

Chang-Keon Ryu, Hwan-Gue, Hyong-Jun Kim. (March 9-12, 2009) Honolulu, Hawaii, USA, A detecting and tracing algorithm for unauthorized internet-news plagiarism using spatio temporal document evolution model”, in Proceedings of the 2009 ACM Symposium on Applied Computing (SAC).

Michael Granitzer, Vedran Sabol, Werner Klieber, Wolfgang Kienreich. (4-8 September 2006), Krakow, Poland. (“Plagiarism Detection in Large Sets of Press Agency News Articles”, in 17th International Workshop on Database and Expert Systems Applications (DEXA 2006), available at (https://scihub.tw/10.1109/DEXA.2006.112).

Kanokorn Trakultaweekoon, Pornpimon Palingoon, Santipong Thaiprayoon. (2019) “Design and Development of a Plagiarism Corpus in Thai for Plagiarism Detection”, 11th International Conference on Knowledge and Systems Engineering (KSE).

Mahmoud, A., Zrigui, A., & Zrigui, M. (2017). A text semantic similarity approach for Arabic paraphrase detection. In International conference on computational linguistics and intelligent text processing, (pp. 338–349). Cham: Springer.

Khorsi, A., Cherroun, H., & Schwab, D. (2018). 2L-APD: A two-level plagiarism detection system for Arabic documents. Cybernetics and Information Technologies, 18(1), 124–138.

Shanta Phani, S. Lahiri, A. Biswas (2015) “Authorship Attribution in Bengali Language”, published in ICON 2015.

wikipedia. (n.d.) Retrieved from wikipedia https://en.wikipedia.org/wiki/Plagiarism

kaggle. (n.d.) Retrived from Kaggle https://www.kaggle.com/zshujon/40k-bangla-newspaper-article

plagiarism.org. (n.d.) Retrieved from plagiarism.org https://www.plagiarism.org/article/what-is-plagiarism