Bangla Article Similarity and recommendation using Cosine Similarity

Abstract

Automated Bangla article similarity is basically finding the most relevant date that resembles with search text.. Despite several comprehensive textual datasets are available for different languages, a few small datasets are curated on Bangla language. As a result, a few works address Bangla document recommendation problem, and due to the lack of enough data. In this work, we created a large dataset of Bangla articles from **The Daliy Ittefaq**, which contains around 70k articles. This huge diverse dataset helps us to create a system by utilizing TF-IDF features, which finally gives recommendation.

Methodology

We have used various features in this project. They would be clarified by following segments:

Dataset

For dataset, We have used the articles from famous Bengali newspaper “The Daily Ittefaq”.Almost 54K documents have been retrieved from the website of the newspaper for the years 2019 and 2020.Almost 9,00,000 sentences are present in the corpus.

Summary

#Total Article count ==========> 53571

#unique unigram ==========> 252073

#Total Unigram ==========>13326446

#unique bigram ==========> 3718409

#Total Bigram ==========> 13579033

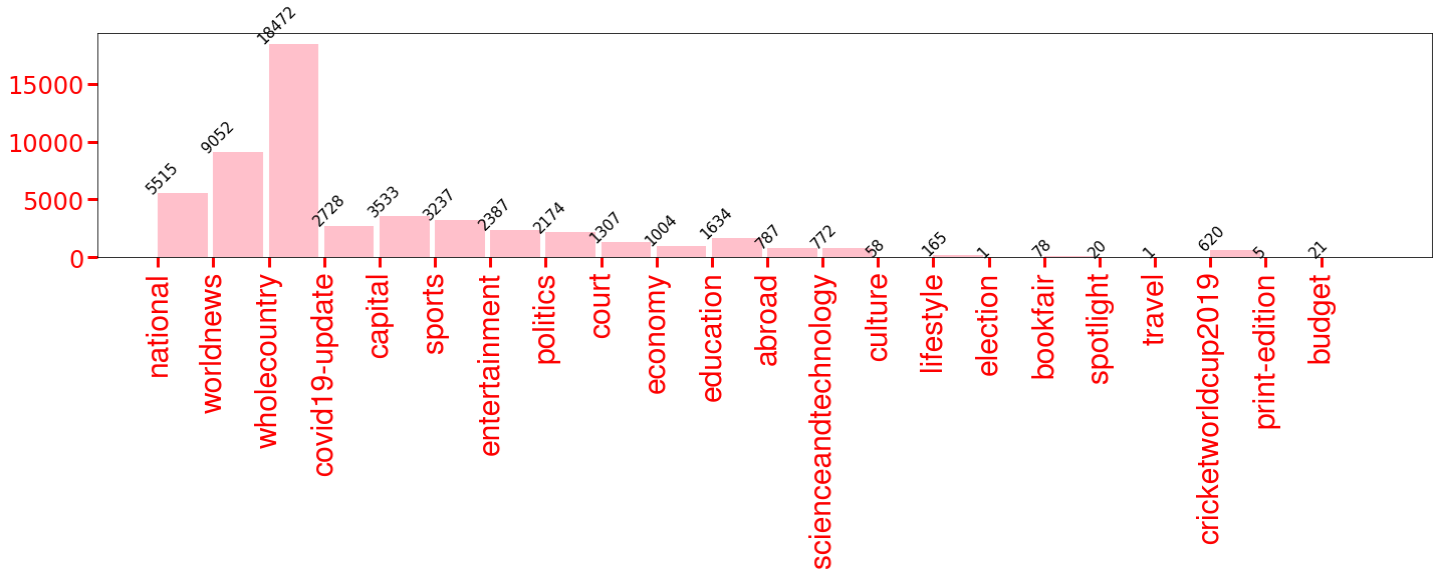
#unique trigram ==========> 8218523

#Total Trigram ==========> 24030358

Category Summary: **Total 22 Categories**

Type	News Count	Type (Bengali Version)

'national':	5515,	জাতীয়
'worldnews':	9052,	আন্তর্জাতিক
'wholecountry':	18472,	সারাদেশ
'covid19-update':	2728,	কোভিড-১৯
'capital':	3533,	রাজধানী
'sports':	3237,	খেলা
'entertainment':	2387,	বিনোদন
'politics':	2174,	রাজনীতি
'court':	1307,	কোর্ট
'economy':	1004,	অর্থনীতি
'education':	1634,	শিক্ষা
'abroad':	787,	বিদেশ
'Scienceandtechnology':	772,	বিজ্ঞান ও টেক
'culture':	58,	সংস্কৃতি
'lifestyle':	165,	জীবনযাপন
'bookfair':	78,	বইমেলা
'cricketworldcup2019':	620	ক্রিকেট বিশ্বকাপ -১৯



So, the methodology is consisted of following procedures:

A. Preprocessing

Preprocessing is very important in terms of raw data is concerned and especially in Bengali language where different punctuation marks are there.

a. Punctuation Removal

First of all, I have ensured the proper cleaning of each document by cleaning the punctuation marks. Again many Unicode were not detected, so they were needed to be cleaned as well.

b. Duplicate Sentence Removal

There were many duplicate sentences in the corpus as well. So I have made sure that they unique sentences will only be there clearing the duplicate ones.

c. Stop words Removal

There are many stop words in Bengali languages like “ও” , “এবং” ,”আর” etc.These stopwords are identified by computing frequency based unigrams.The most frequent words were the stop words so many stopwords were identified from the frequency based unigrams and from Bengali language point of view.

d .Stemming

There are several words in different formations like শহর, শহরে,শহরের etc. They basically denote the same meaning শহর but with the inclusion of terms like ‘s’, ‘es’ it shows the different form. To remove redundancy, the conversion of several formations into an actual word is also needed.

e. Tokenization

Word tokenization is done here for each document to compute the tf-idfs of each word of each document which is described in the coming sections.

f.Small documents removal

Article of smaller lengths are removed in this case.

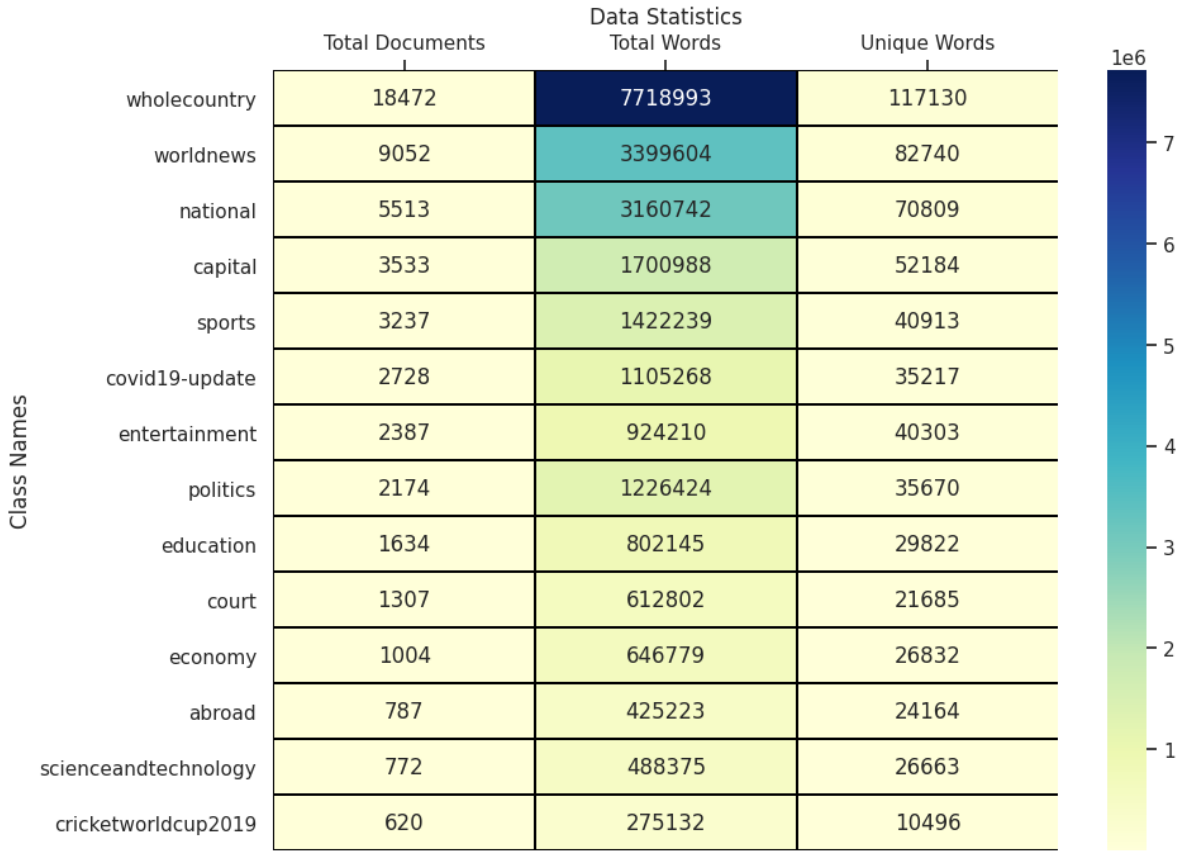


Fig:After pre-processing dataset

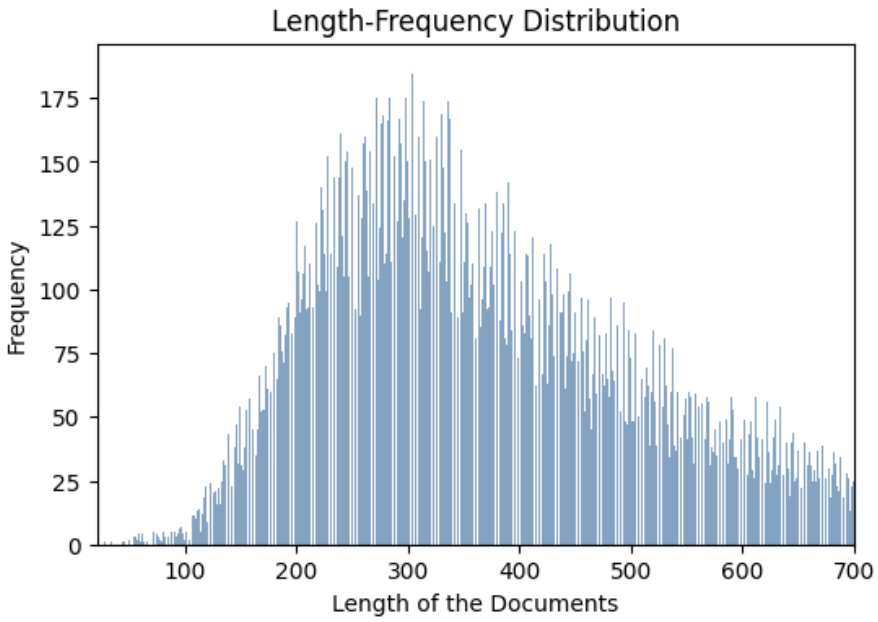


Fig:Lengths of documents

B. Bangla Article Similarity and recommendation using Cosine Similarity

a.TF-IDF

TF-IDF is a measurement that symbolizes how important a word is to a document in a association of documents. Two metrics are multiplied in this regard. They are the term frequency and the inverse document frequency. Term frequency is how many times a word appears in a document divided by the length of the document in terms of word count and the inverse document frequency is the number of total documents divided by the number of documents that the word has appeared in.

$$\text{TF*IDF}= \frac{No\ of\ perticualr\ term\ in\ that\ doccument}{Total\ no\ nof\ term\ in\ that\ doccument} * \frac{Total\ No\ of\ doccument}{No\ of\ doccument\ where\ the\ perticular\ word\ appears}$$

…….(1)

Fig: TF-IDF

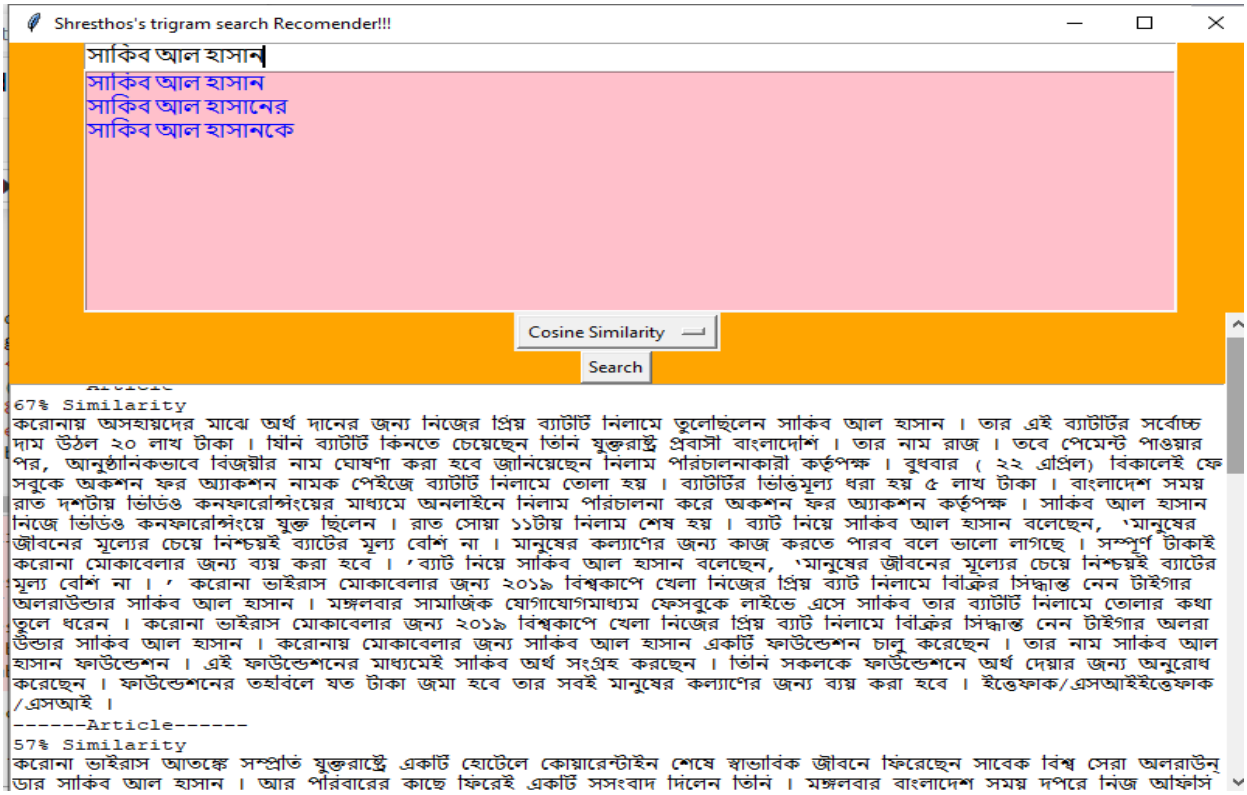
b. Cosine Similarity

Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis.

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

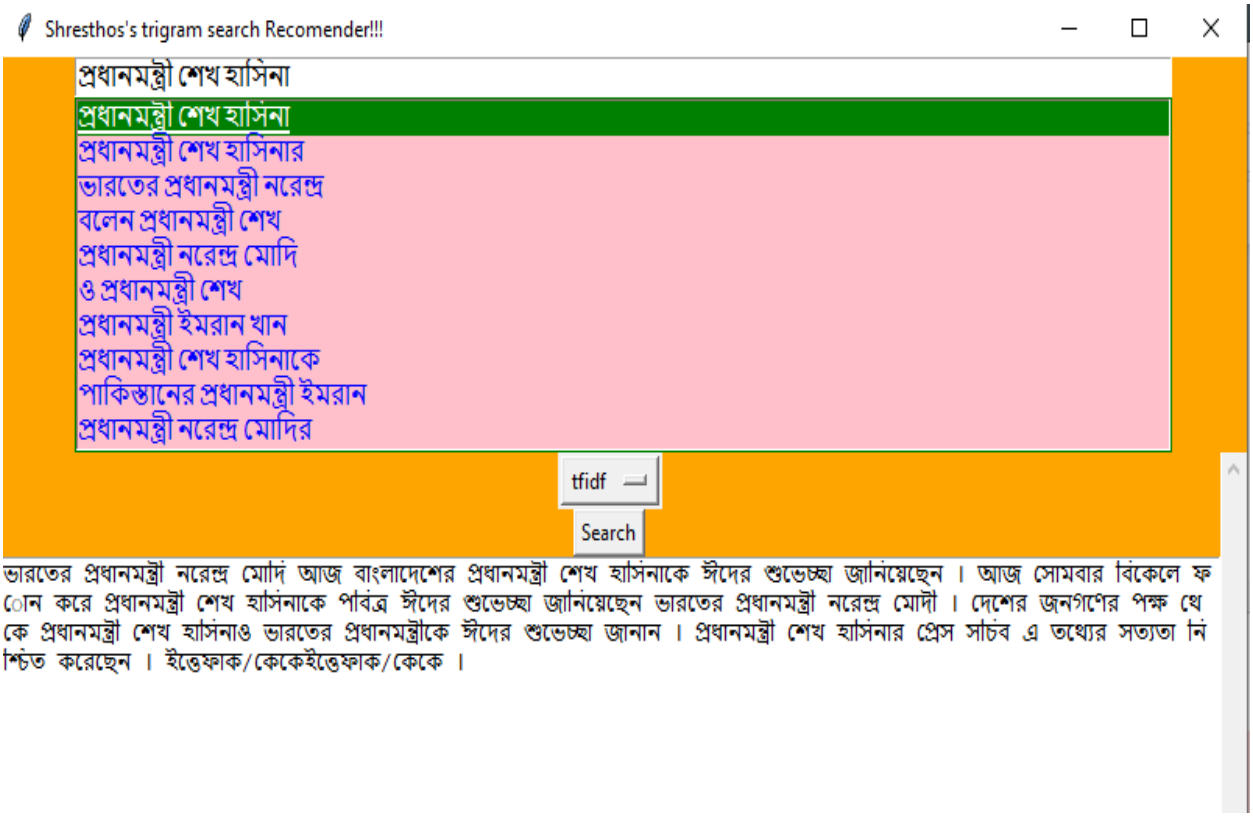
…….(2)

Using the search text the similarity with articles are measured and top 5 articles are recommended.



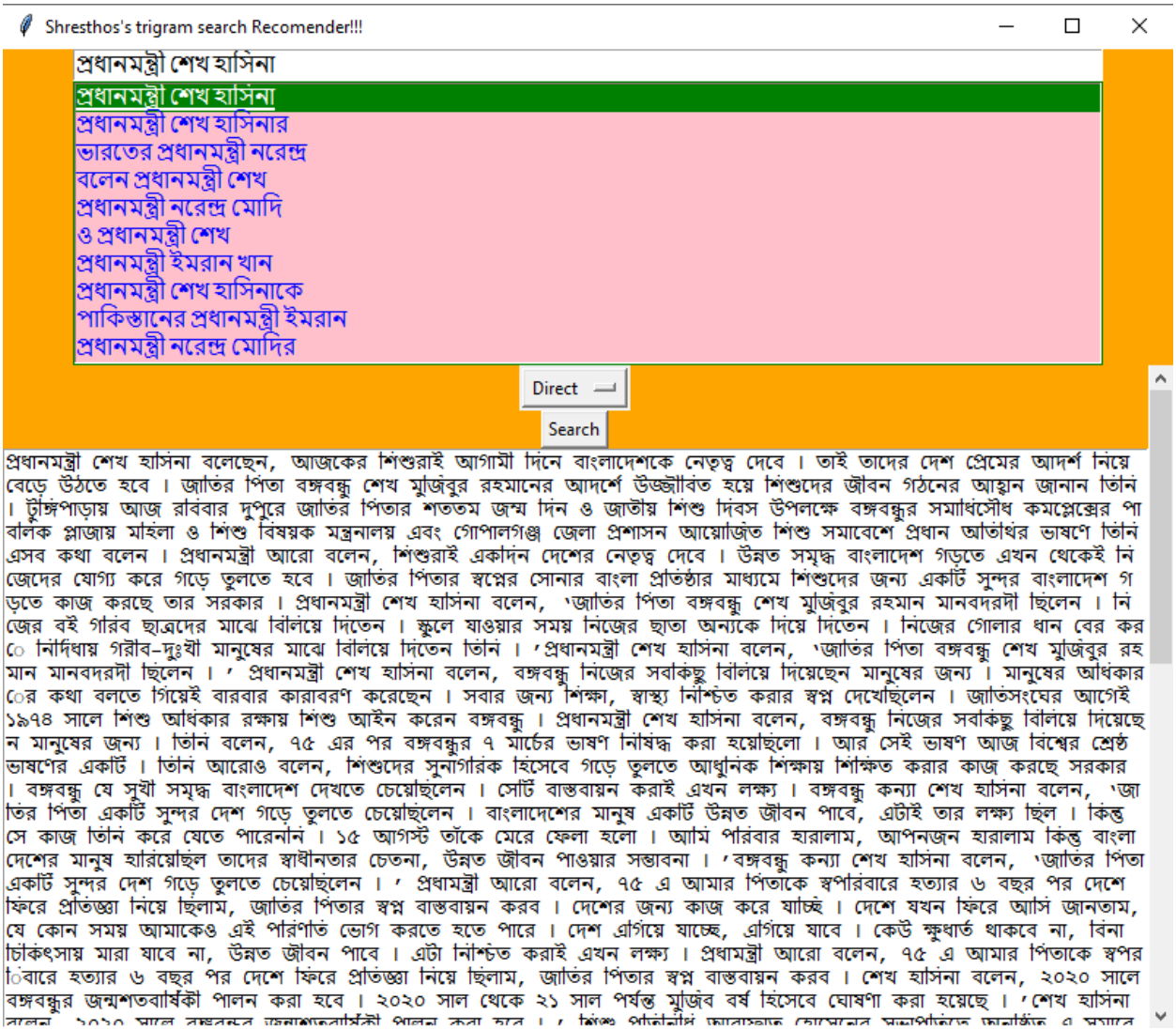
C. Bangla Article Similarity and recommendation using TF-IDF

In this method ,the search strings are tokenized and for those tokens the tfidf values are calculated and the highest valued article is returned.



D. Bangla Article Similarity and recommendation using Direct Search

In this method, the searched string is searched in all the documents and the document that has the highest frequency of the search text is returned.The following figure shows it.



প্রধানমন্ত্রী শেখ হাসিনা বলেছেন, আজকের শিশুরাই আগামী দিনে বাংলাদেশকে নেতৃত্ব দেবে । তাই তাদের দেশ প্রেমের আদর্শ নিয়ে বেড়ে উঠতে হবে । জাতির পিতা বঙ্গবন্ধু শেখ মুজিবুর রহমানের আদর্শে উজ্জীবিত হয়ে শিশুদের জীবন গঠনের আহ্বান জানান তিনি । টুঙ্গিপাড়ায় আজ রবিবার দুপুরে জাতির পিতার শততম জন্ম দিন ও জাতীয় শিশু দিবস উপলক্ষে বঙ্গবন্ধুর সমাধিসৌধ কমপ্লেক্সের পাবলিক প্লাজায় মহিলা ও শিশু বিষয়ক মন্ত্রনালয় এবং গোপালগঞ্জ জেলা প্রশাসন আয়োজিত শিশু সমাবেশে প্রধান অতিথির ভাষণে তিনি এসব কথা বলেন । প্রধানমন্ত্রী আরো বলেন, শিশুরাই একদিন দেশের নেতৃত্ব দেবে । উন্নত সমৃদ্ধ বাংলাদেশ গড়তে এখন থেকেই নিজেদের যোগ্য করে গড়ে তুলতে হবে । জাতির পিতার স্বপ্নের সোনার বাংলা প্রতিষ্ঠার মাধ্যমে শিশুদের জন্য একটি সুন্দর বাংলাদেশ গড়তে কাজ করছে তার সরকার । প্রধানমন্ত্রী শেখ হাসিনা বলেন, 'জাতির পিতা বঙ্গবন্ধু শেখ মুজিবুর রহমান মানবদরদী ছিলেন । নিজের বই গরিব ছাত্রদের মাঝে বিলিয়ে দিতেন । স্কুলে যাওয়ার সময় নিজের ছাতা অন্যকে দিয়ে দিতেন । নিজের গোলার ধান বের করে নির্দিধায় গরীব-দুঃখী মানুষের মাঝে বিলিয়ে দিতেন তিনি । 'প্রধানমন্ত্রী শেখ হাসিনা বলেন, 'জাতির পিতা বঙ্গবন্ধু শেখ মুজিবুর রহমান মানবদরদী ছিলেন । ' প্রধানমন্ত্রী শেখ হাসিনা বলেন, বঙ্গবন্ধু নিজের সবকিছু বিলিয়ে দিয়েছেন মানুষের জন্য । মানুষের অধিকার কথা বলতে গিয়েই বারবার কারাবরণ করেছেন । সবার জন্য শিক্ষা, স্বাস্থ্য নিশ্চিত করার স্বপ্ন দেখেছিলেন । জাতিসংঘের আগেই ১৯৭৪ সালে শিশু অধিকার রক্ষায় শিশু আইন করেন বঙ্গবন্ধু । প্রধানমন্ত্রী শেখ হাসিনা বলেন, বঙ্গবন্ধু নিজের সবকিছু বিলিয়ে দিয়েছেন মানুষের জন্য । তিনি বলেন, ৭৫ এর পর বঙ্গবন্ধুর ৭ মার্চের ভাষণ নিষিদ্ধ করা হয়েছিলো । আর সেই ভাষণ আজ বিশ্বের শ্রেষ্ঠ ভাষণের একটি । তিনি আরোও বলেন, শিশুদের সুনাগরিক হিসেবে গড়ে তুলতে আধুনিক শিক্ষায় শিক্ষিত করার কাজ করছে সরকার । বঙ্গবন্ধু যে সুখী সমৃদ্ধ বাংলাদেশ দেখতে চেয়েছিলেন । সেটি বাস্তবায়ন করাই এখন লক্ষ্য । বঙ্গবন্ধু কন্যা শেখ হাসিনা বলেন, 'জাতির পিতা একটি সুন্দর দেশ গড়ে তুলতে চেয়েছিলেন । বাংলাদেশের মানুষ একটি উন্নত জীবন পাবে, এটাই তার লক্ষ্য ছিল । কিন্তু সে কাজ তিনি করে যেতে পারেননি । ১৫ আগস্ট তাঁকে মেরে ফেলা হলো । আমি পরিবার হারালাম, আপনজন হারালাম কিন্তু বাংলাদেশের মানুষ হারিয়েছিল তাদের স্বাধীনতার চেতনা, উন্নত জীবন পাওয়ার সম্ভাবনা । 'বঙ্গবন্ধু কন্যা শেখ হাসিনা বলেন, 'জাতির পিতা একটি সুন্দর দেশ গড়ে তুলতে চেয়েছিলেন । ' প্রধামন্ত্রী আরো বলেন, ৭৫ এ আমার পিতাকে স্বপরিবারে হত্যার ৬ বছর পর দেশে ফিরে প্রতিজ্ঞা নিয়ে ছিলাম, জাতির পিতার স্বপ্ন বাস্তবায়ন করব । দেশের জন্য কাজ করে যাচ্ছি । দেশে যখন ফিরে আসি জানতাম, যে কোন সময় আমাকেও এই পরিণতি ভোগ করতে হতে পারে । দেশ এগিয়ে যাচ্ছে, এগিয়ে যাবে । কেউ ক্ষুধার্ত থাকবে না, বিনা চিকিৎসায় মারা যাবে না, উন্নত জীবন পাবে । এটা নিশ্চিত করাই এখন লক্ষ্য । প্রধামন্ত্রী আরো বলেন, ৭৫ এ আমার পিতাকে স্বপরিবারে হত্যার ৬ বছর পর দেশে ফিরে প্রতিজ্ঞা নিয়ে ছিলাম, জাতির পিতার স্বপ্ন বাস্তবায়ন করব । শেখ হাসিনা বলেন, ২০২০ সালে বঙ্গবন্ধুর জন্মশতবার্ষিকী পালন করা হবে । ২০২০ সাল থেকে ২১ সাল পর্যন্ত মুজিব বর্ষ হিসেবে ঘোষণা করা হয়েছে । 'শেখ হাসিনা বলেন, ২০২০ সালে বঙ্গবন্ধুর জন্মশতবার্ষিকী পালন করা হবে । ' শিশু পরিবারের আবাসিক স্কুলসমূহের সমন্বয়কারী অনুষ্ঠিত এ সম্মাবে