Bangla sequence prediction using LSTMs

Abstract

Automated Bangla sequence prediction is basically finding out what data is most likely to be there after the input data. 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 sequence prediction problem, and due to the lack of enough training data, these approaches could not able to learn sophisticated supervised learning model. 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 lstm model by utilizing word embeddings,TF-IDF features,which finally predicts Sequences.

Methodology

Word embedding are most interesting features in terms of natural language processing is concerned.We have used word embeddings as features in this project.For deep learning model **l lstm(Long Short Term Memory).**The methodology is consisted of some following proceddings:

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

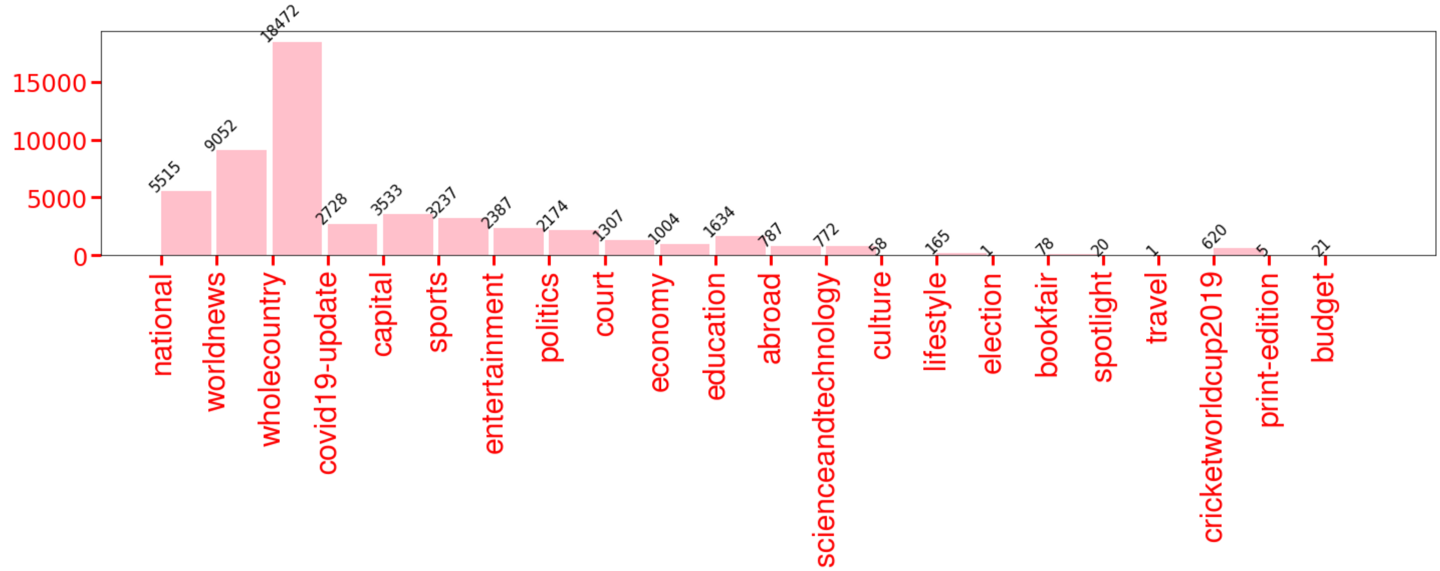
#Total Bigram ==============> 13579033

#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

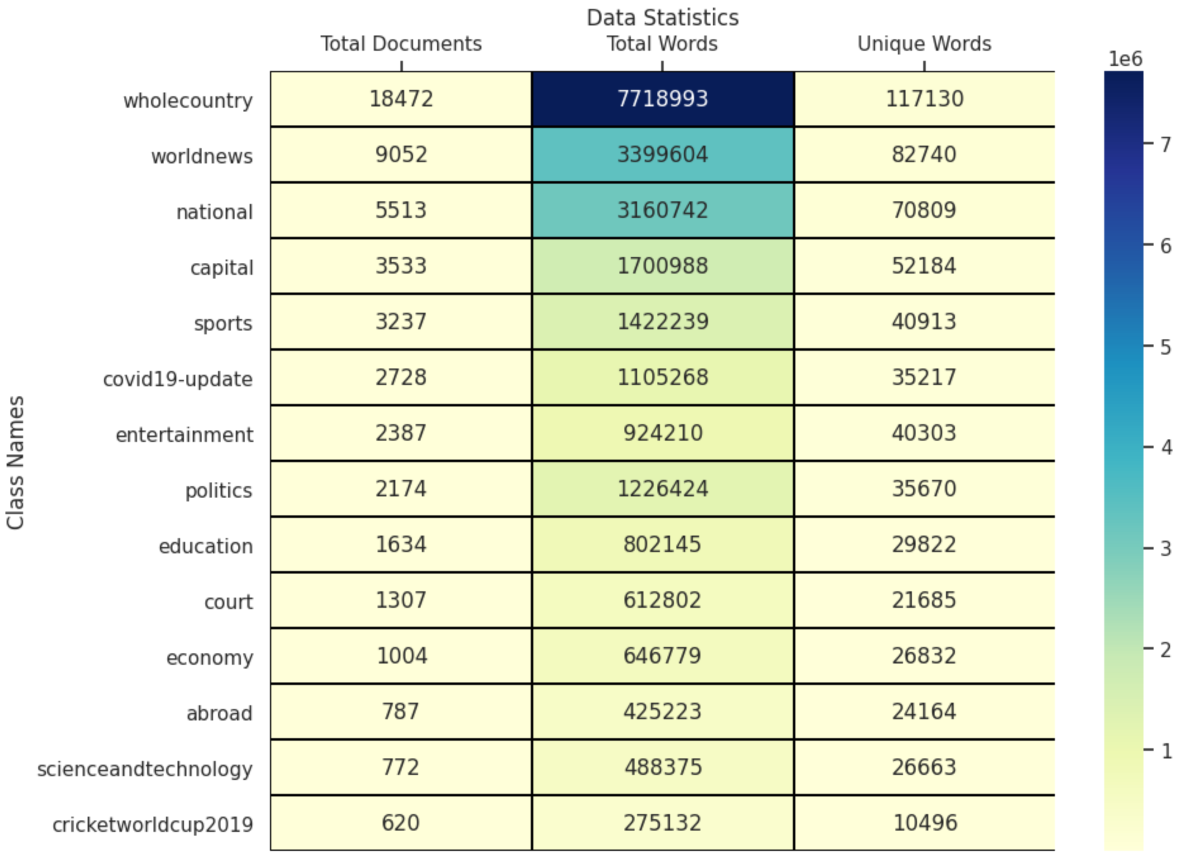
There are several words in different formations like শহর, শহরে,শহরের etc. They basically denote the same meaning শহর but with the inclusion of terms like ‘s’, ‘es’ it shows the different form. To remove redundancy, the conversion of several formations into an actual word is also needed.

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.



	Total Documents	Total Words	Unique Words
wholecountry	18472	7718993	117130
worldnews	9052	3399604	82740
national	5513	3160742	70809
capital	3533	1700988	52184
sports	3237	1422239	40913
covid19-update	2728	1105268	35217
entertainment	2387	924210	40303
politics	2174	1226424	35670
education	1634	802145	29822
court	1307	612802	21685
economy	1004	646779	26832
abroad	787	425223	24164
scienceandtechnology	772	488375	26663
cricketworldcup2019	620	275132	10496

Fig:After pre-processing dataset

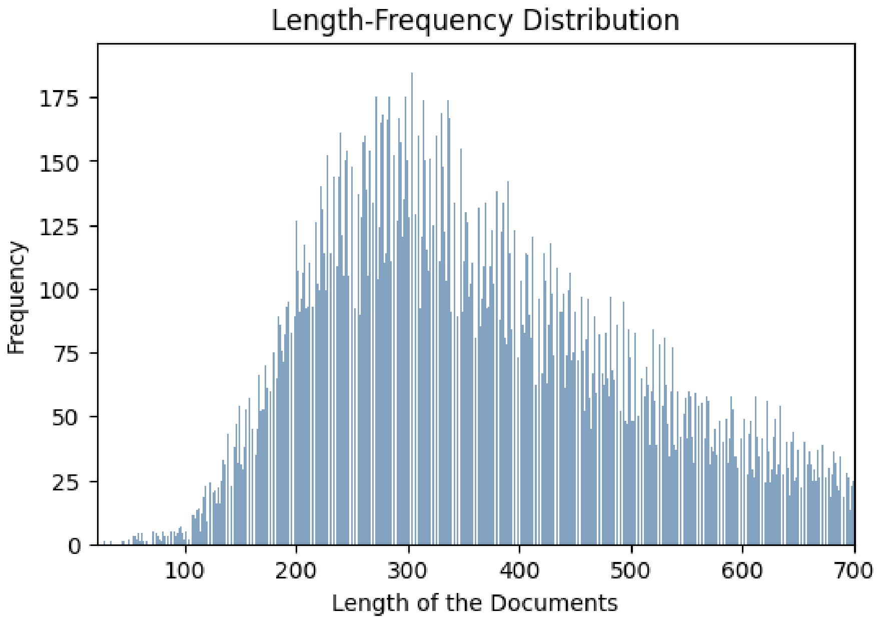


Fig:Lengths of documents

B. Bangla sequence prediction using LSTMs

Word Embedding

A word embedding is a learned representation for text where words that have the same meaning have a similar representation. It is this approach to representing words and documents that may be considered one of the key breakthroughs of deep learning on challenging natural language processing problems. For this an embedding layer is used.

LSTM

Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997), and were refined and popularized by many people in following work.1 They work tremendously well on a large variety of problems, and are now widely used.

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn

LSTM learns sequences quite well which will result in the best sequence prediction.

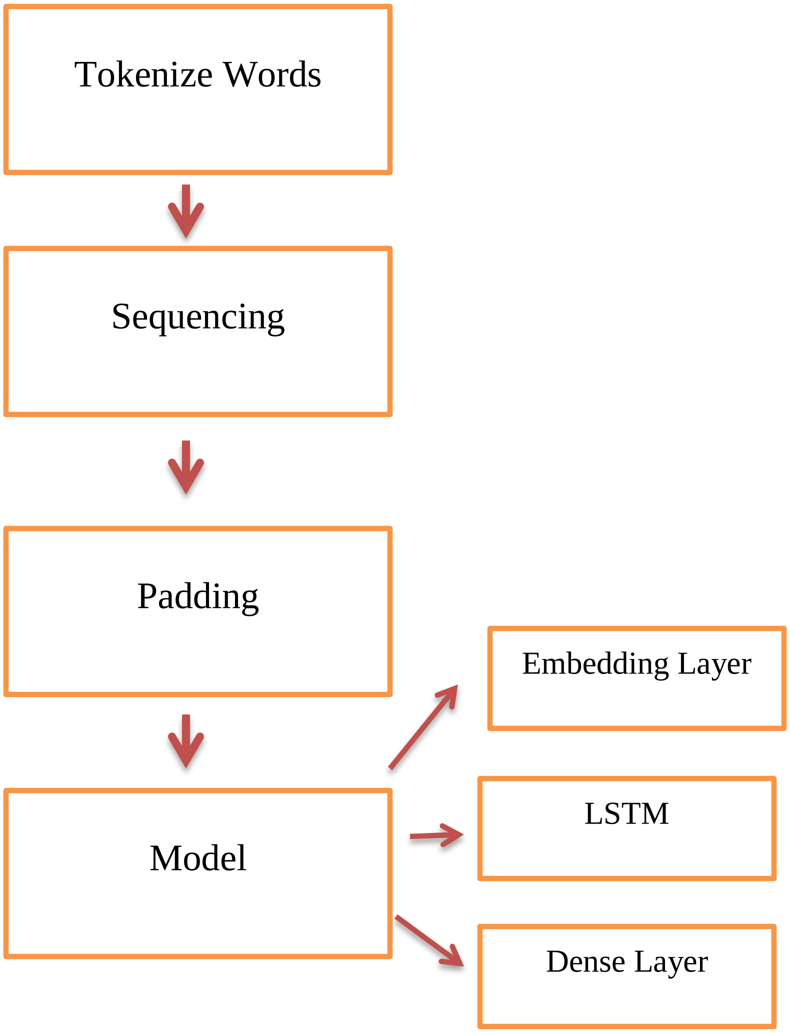
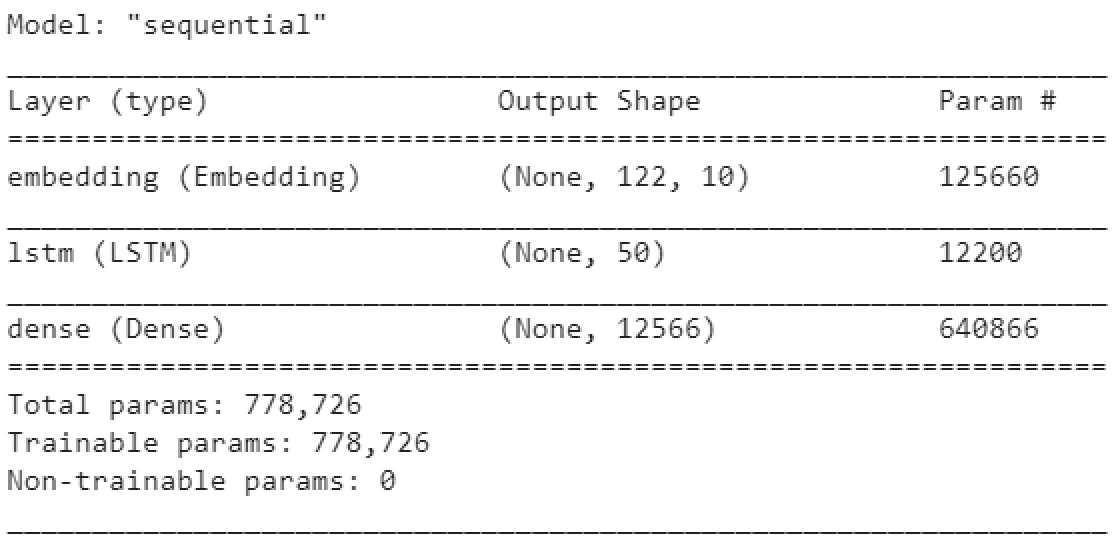


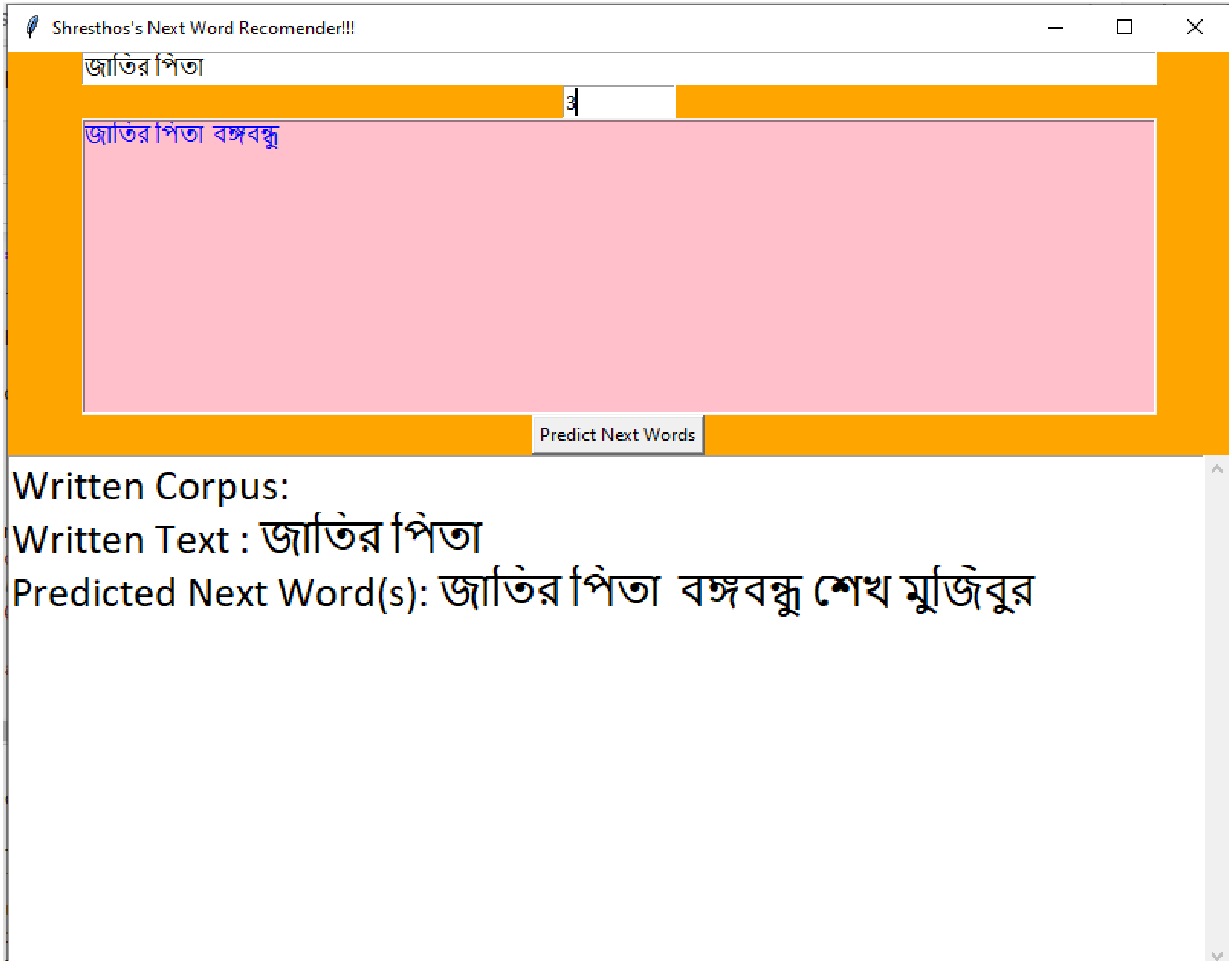
Fig: Bangla sequence prediction using LSTMs

Model used:



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
embedding (Embedding)        (None, 122, 10)           125660    
_________________________________________________________________
lstm (LSTM)                  (None, 50)                12200     
_________________________________________________________________
dense (Dense)                (None, 12566)             640866    
=================================================================
Total params: 778,726
Trainable params: 778,726
Non-trainable params: 0
_________________________________________________________________
```

C. Experimental results and evaluation



জাতির পিতা

3

জাতির পিতা বঙ্গবন্ধু

Written Corpus:

Written Text : জাতির পিতা

Predicted Next Word(s): জাতির পিতা বঙ্গবন্ধু শেখ মুজিবুর

D. Application UI

Computer based software is implemented utilizing proposed method. Following figure shows that. For the UI ,Tkinter library is used for python.