Bangla Fake News Detection

Protik Bose Pranto (1505044)
Soumit Kanti Saha (1505047)

Dataset

BanFakeNews: A Dataset for Detecting Fake News in Bangla

Md Zobaer Hossain[†♣], Md Ashraful Rahman[†♣], Md Saiful Islam[♣], Sudipta Kar[♠]
[♣] Shahjalal University of Science and Technology, Sylhet, Bangladesh
{zobaer37, ashraful54}@student.sust.edu, saiful-cse@sust.edu
[♠] University of Houston, Texas, USA
skar3@uh.edu

```
merged['label'].value_counts()
```

```
1    49192
0     3425
Name: label, dtype: int64
```

1 -> Authentic News
0 -> Fake News

Dataset

	articleID	domain	date	category	source	relation	headline	content	label	F-type
0	15060	banglanews24.com	2018-09-26 01:36:42	National	NaN	NaN	'জাতীয় ঐক্য'র বিরুদ্ধে 'প্রত্যয়' ঘোষণা ১৪ দলের	ঢাকা: নবগঠিত 'বৃহত্তর জাতীয় ঐক্য'র বিরুদ্ধে '...	1.0	NaN
1	24764	somoynews.tv	2018-09-28 14:12:47	International	NaN	NaN	ভয়ঙ্কর মানুশখেকো এই নারী, ঘরে মিলল মানুষের ত্ব...	মানুষের মাংস কোন মানুষ খায়, একথা শুনলেও তো কেম...	1.0	NaN
2	21857	bangla.thereport24.com	2018-09-29 13:28:05	Crime	NaN	NaN	ফার্মগেটে বাসের ধাক্কায় কৃষি কর্মকর্তা নিহত	দ্য রিপোর্ট প্রতিবেদক : রাজধানীর ফার্মগেট এলাক...	1.0	NaN
3	1132	jugantor.com	2018-09-21 11:55:08	Editorial	Reporter	Related	প্রধানমন্ত্রী ও বিরোধীদলীয় নেত্রীর রুদ্ধদ্বার ...	প্রধানমন্ত্রী শেখ হাসিনার সঙ্গে বৈঠক করেছেন বি...	0.0	NaN
4	16813	jagonews24.com	2018-09-25 18:11:12	National	NaN	NaN	স্মার্টওয়াচ না পেয়ে স্কুলছাত্রের আত্মহত্যা	স্মার্টওয়াচ না পেয়ে অরবিন্দু রায় (১৪) নামের ন...	1.0	NaN

Dataset

- articleID : ID of the news
- domain : News publisher's site name
- date : Published Date
- category : Category of the news
- source : Source of the news. (One who can verify the claim of the news)
- relation : Related or Unrelated. Related if headline matches with content's claim otherwise it is labeled as Unrelated
- headline : Headline of the news
- content : Article or body of the news
- label : 1 or 0 . '1' for authentic '0' for fake
- F-type : Type of fake news (Clickbait, Satire, Fake(Misleading or False Context))

Preprocessing

- Unique Content Handling
- Handling Null Values

```
merged.isnull().sum()
```

```
articleID        0
domain           0
date             0
category         0
source       49977
relation     49977
headline         0
content          0
label            0
F-type       57179
dtype: int64
```

```
merged.drop(['source'], axis=1, inplace=True)
merged.drop(['relation'], axis=1, inplace=True)
merged.drop(['F-type'], axis=1, inplace=True)
```

```
# remove articleID
merged.drop(['articleID'], axis=1, inplace=True)
```

Bangla Text Preprocessing

BanglaKit Bengali Stemmer

Bengali Stemmer passing

A stemmer is a light-weight approach to find root words, avoiding expensive morphological analysis. The *BanglaKit Stemmer* implements a stepwise approach to removing inflections from Bengali Words [1].

Work is in progress with the algorithm of the stemmer, the implementations may vary significantly from version to version.

Algorithms

Rafi Kamal's Stemmer

Originally Developed by Rafi Kamal. Ported to Python.:

```
from bengali_stemmer.rafikamal2014 import RafiStemmer
stemmer = RafiStemmer()
stemmer.stem_word('বাংলায়')
```

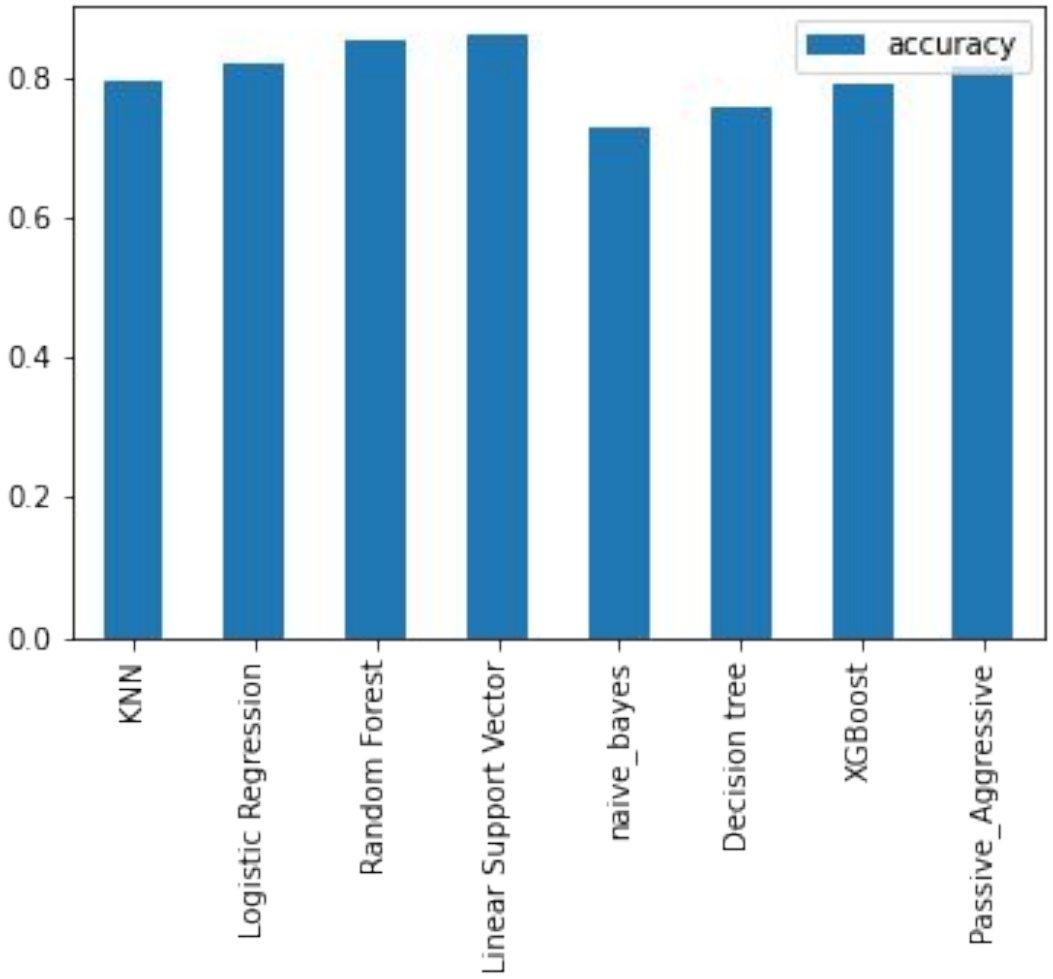
Bangla Text Preprocessing

- Remove HTML Tag
- Remove Hyperlinks
- Remove Punctuations
- Remove Stopwords
- Remove Foreign words
- Remove Numbers
- Stemming

ML Algorithms

- Almost 7000 data (equal number of positive and negative data)
- Tf-Idf vectorization
- 20% Test data
- Only “Text” and “Label” columns are considered

ML Algorithm	Accuracy
KNN	79.42 %
Logistic Regression	82.04 %
Random Forest	**85.38 %**
Linear Support Vector	**85.96 %**
Naive Bayes	72.87 %
Decision Tree	75.85 %
XGBoost	79.20 %
Passive Aggressive Classifier	81.67 %


accuracy
0.8
0.6
0.4
0.2
0.0
KNN
Logistic Regression
Random Forest
Linear Support Vector
naive_bayes
Decision tree
XGBoost
Passive_Aggressive

Upsampling

- 20% Test data
- positive data in training: 39323
- negative data in training: 2770
- positive data in test: 9831
- negative data in test: 693

<u>After upsampling</u> -

- Positive train data = 39323, Negative Data = 39323
- Train data = 78646, Test data = 10524

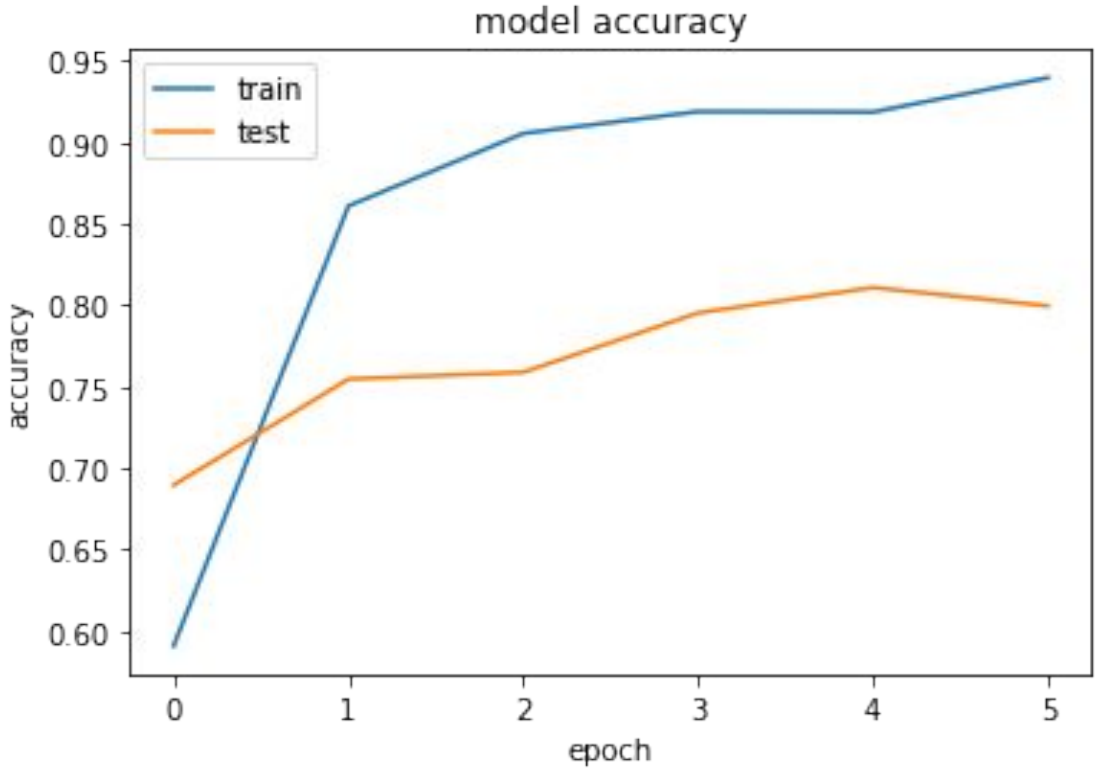
Deep Learning Model

- Recurrent Neural Network Architecture
- Maximum Features 8000 (first 8000 words)
- Maximum Sentence Length 2000
- the vocabulary index based on word frequency (fit_on_texts)
- each text in texts to a sequence of integers (texts_to_sequences)
- Pad sequences

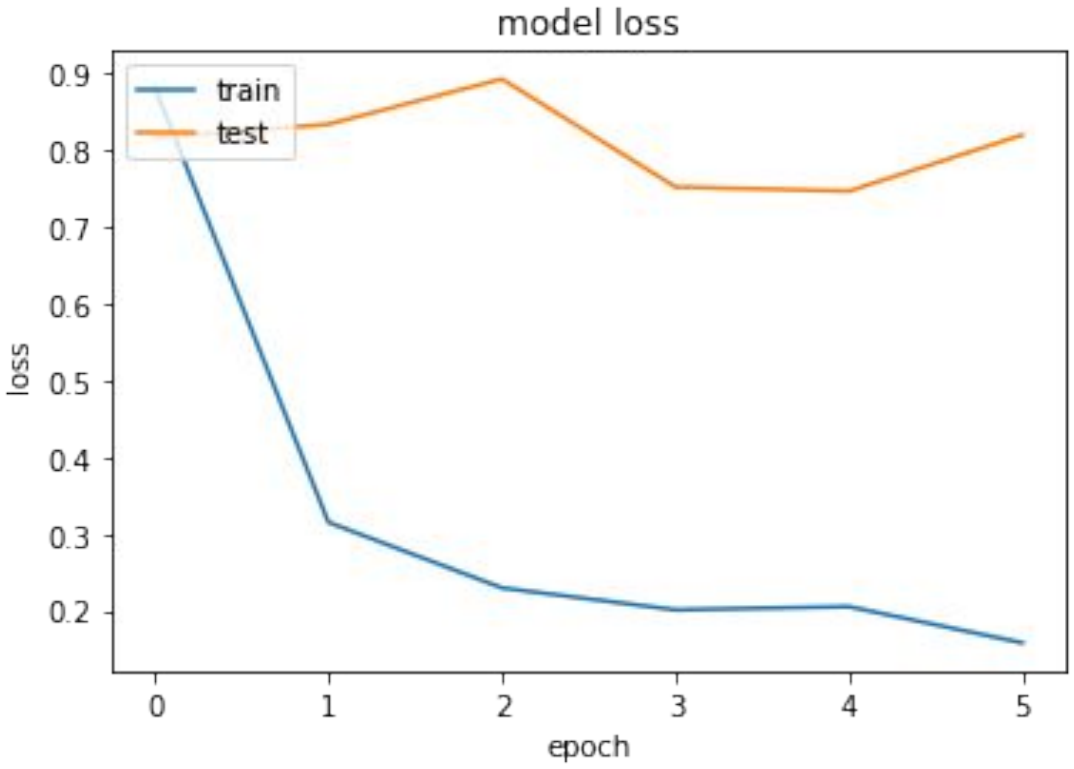
LSTM

- Max_features = 8000, embed_dim = 128
- Embedding (Max_features, embed_dim,input_length = 2000)
- SpatialDropout1D (0.4)
- LSTM (256, dropout=0.4, recurrent_dropout=0.4)
- Dense (2,activation='softmax')
- loss = 'categorical_crossentropy', optimizer='adam',metrics = ['accuracy']
- Class_weight = {**1**:0.54, **0**:7.6}
- Batch Size = 128

LSTM


model accuracy
train
test
accuracy
epoch
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
1
2
3
4
5

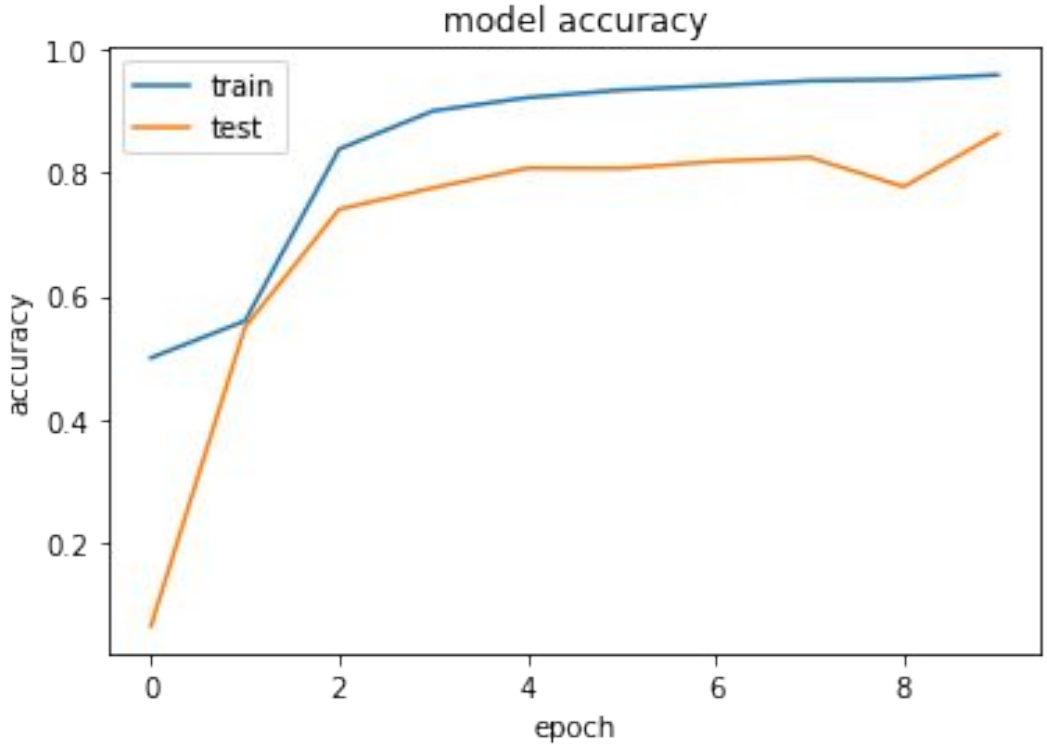
LSTM



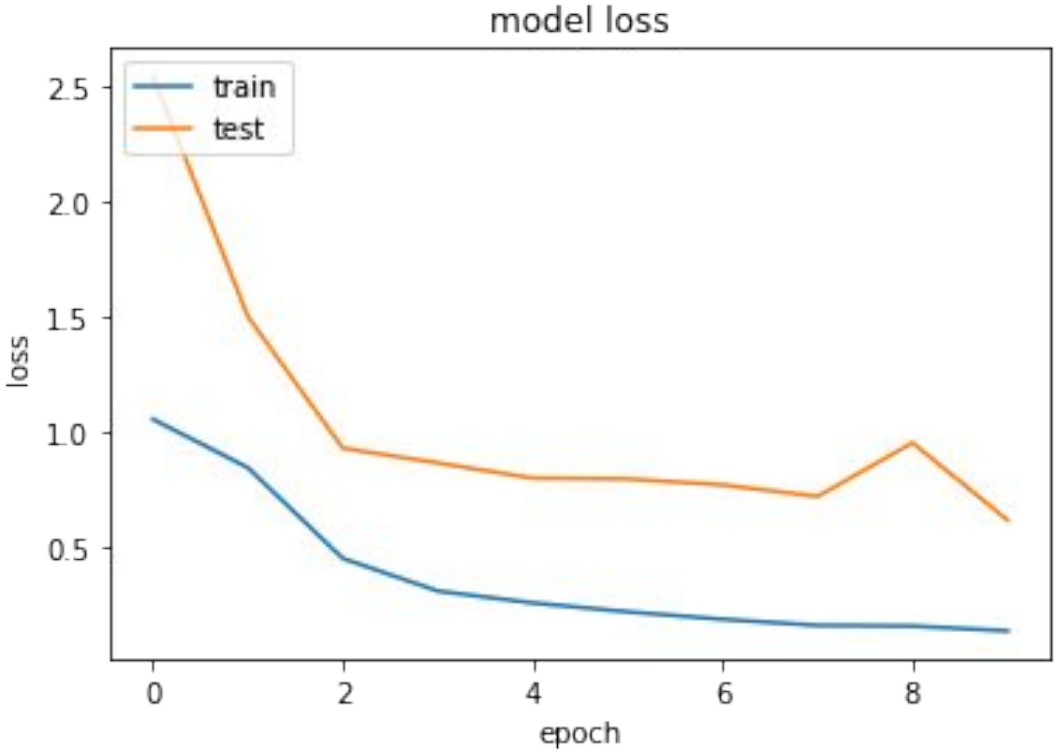
BiLSTM

- Max_features = 8000, embed_dim = 128
- Embedding (Max_features, embed_dim,input_length = 2000)
- Dropout (0.3)
- Bidirectional (LSTM(100))
- Dropout (0.3)
- Dense (32, activation='sigmoid')
- Dense (16, activation='sigmoid')
- Dense (2,activation='softmax')
- loss = 'categorical_crossentropy', optimizer='adam',metrics = ['accuracy']

BiLSTM


model accuracy
train
test
accuracy
epoch
1.0
0.8
0.6
0.4
0.2
0
2
4
6
8

BiLSTM



Any Question?