Bangla Text Classification

Model Used:

In this project I used six models.

1. TF-IDF with Logistic Regression
2. TF-IDF with Random Forest
3. LSTM with Embedding
4. CNN with Embedding
5. Transformers BERT based mode: Multilingual BERT
6. Transformers BERT based mode: Bangla Bert (sagorsarker)

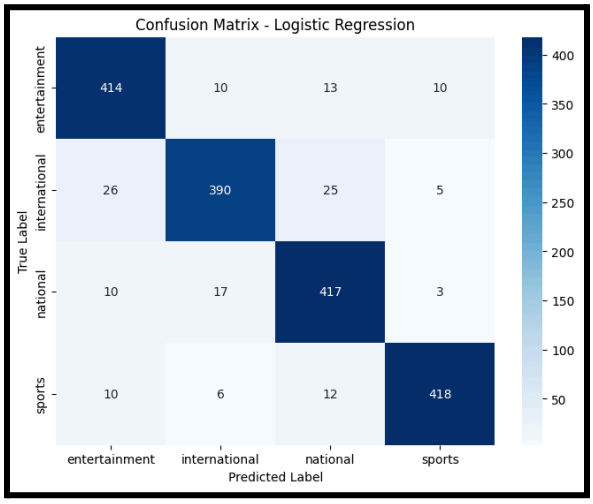
Model's Performance:

Model 1:

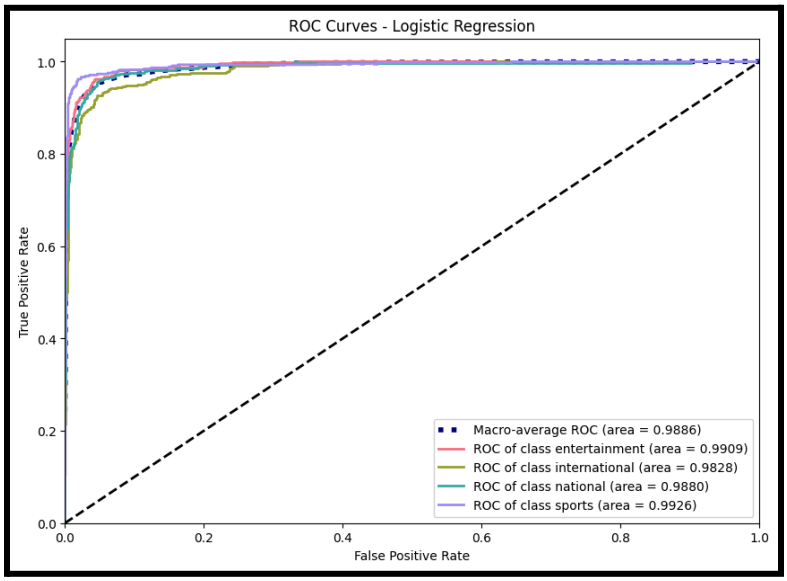
Classification Report

```
--- Logistic Regression Evaluation on TEST SET ---
Test Accuracy: 0.9177

Test Classification Report:
               precision    recall  f1-score   support

entertainment     0.9000    0.9262    0.9129       447
international     0.9220    0.8744    0.8976       446
     national     0.8929    0.9329    0.9125       447
       sports     0.9587    0.9372    0.9478       446

     accuracy                         0.9177      1786
    macro avg     0.9184    0.9177    0.9177      1786
 weighted avg     0.9184    0.9177    0.9177      1786
```

Confusion Matrix:



ROC- Curves:

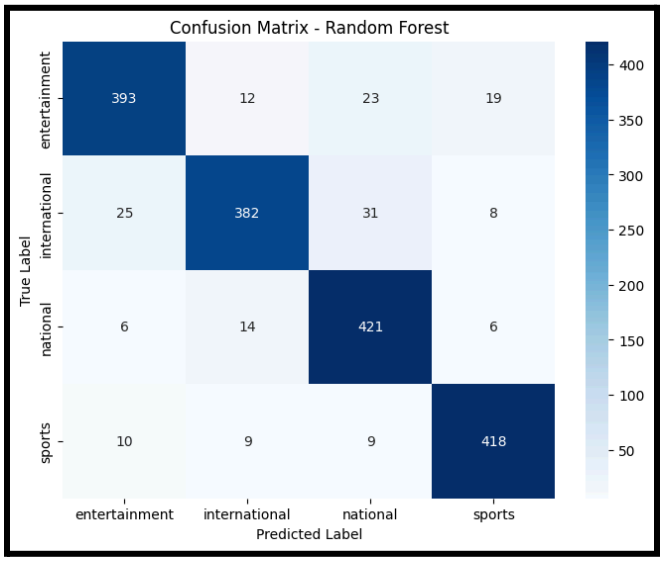


Model 2:

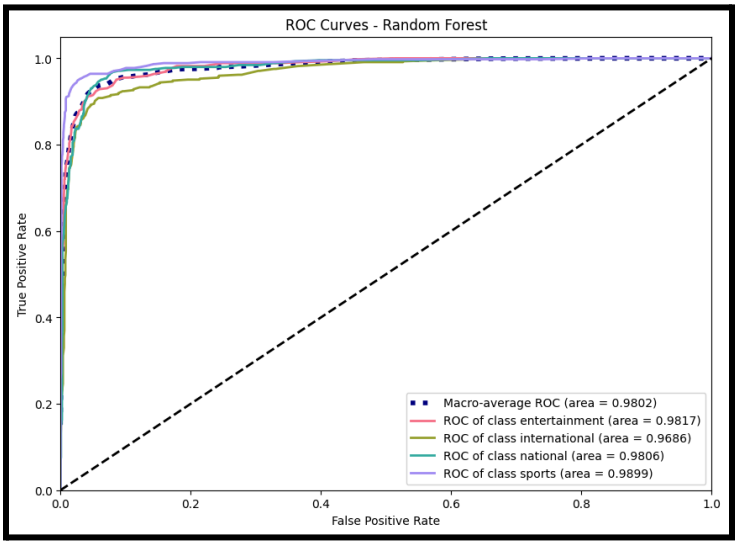
Classification Report:

```
--- Random Forest Evaluation on TEST SET ---
Test Accuracy: 0.9037

Test Classification Report:
               precision    recall  f1-score   support

entertainment     0.9055    0.8792    0.8922       447
international     0.9161    0.8565    0.8853       446
     national     0.8698    0.9418    0.9044       447
       sports     0.9268    0.9372    0.9320       446

     accuracy                         0.9037      1786
    macro avg     0.9046    0.9037    0.9035      1786
 weighted avg     0.9045    0.9037    0.9035      1786
```

Confusion Matrix:



ROC-Curves:

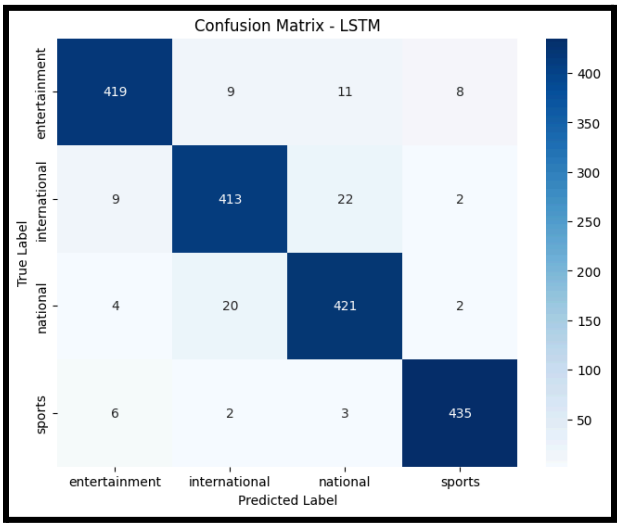


Model 3:

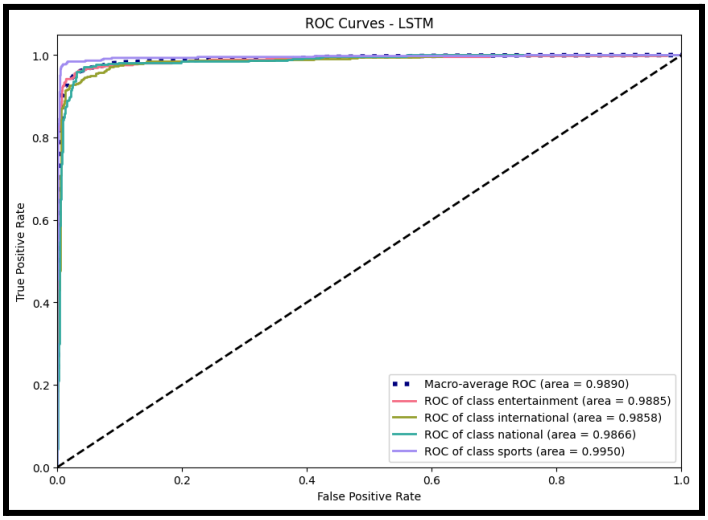
Classification Report:

```
--- LSTM Evaluation on TEST SET ---
Test Accuracy: 0.9451

Test Classification Report:
               precision    recall  f1-score   support

entertainment     0.9566    0.9374    0.9469       447
international     0.9302    0.9260    0.9281       446
     national     0.9212    0.9418    0.9314       447
       sports     0.9732    0.9753    0.9742       446

     accuracy                         0.9451      1786
    macro avg     0.9453    0.9451    0.9452      1786
 weighted avg     0.9453    0.9451    0.9452      1786
```

Confusion Matrix:



ROC-Curve:

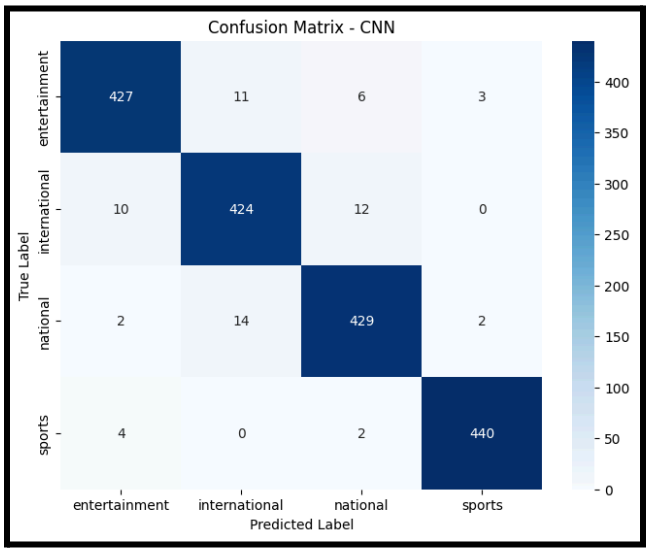


Model 4:

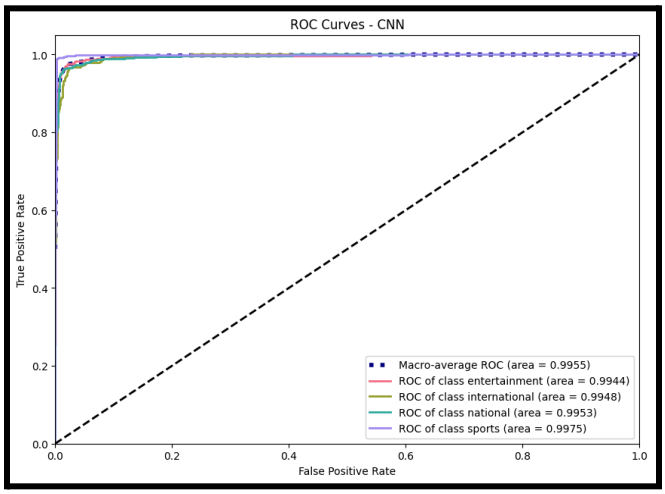
Classification Report:

```
--- CNN Evaluation on TEST SET ---
Test Accuracy: 0.9630

Test Classification Report:
               precision    recall  f1-score   support

entertainment     0.9639    0.9553    0.9596       447
international     0.9443    0.9507    0.9475       446
     national     0.9555    0.9597    0.9576       447
       sports     0.9888    0.9865    0.9877       446

     accuracy                         0.9630      1786
    macro avg     0.9631    0.9631    0.9631      1786
 weighted avg     0.9631    0.9630    0.9631      1786
```

Confusion Matrix:



ROC-Curves:

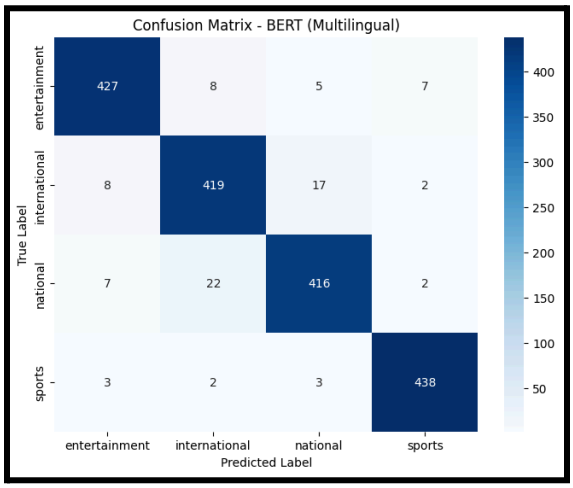


Model 5:

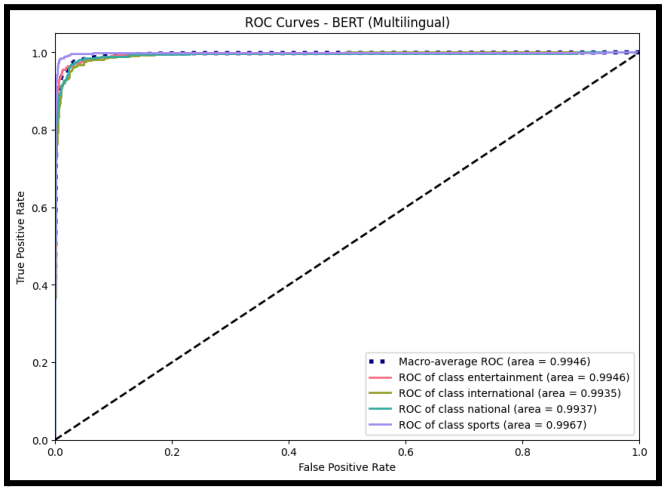
Classification Report:

```
--- BERT (Multilingual) Evaluation on TEST SET ---
Test Accuracy: 0.9518

Test Classification Report:
               precision    recall  f1-score   support

entertainment     0.9596    0.9553    0.9574       447
international     0.9290    0.9395    0.9342       446
     national     0.9433    0.9306    0.9369       447
       sports     0.9755    0.9821    0.9788       446

     accuracy                         0.9518      1786
    macro avg     0.9519    0.9519    0.9518      1786
 weighted avg     0.9519    0.9518    0.9518      1786
```

Confusion Matrix:



ROC-Curves:

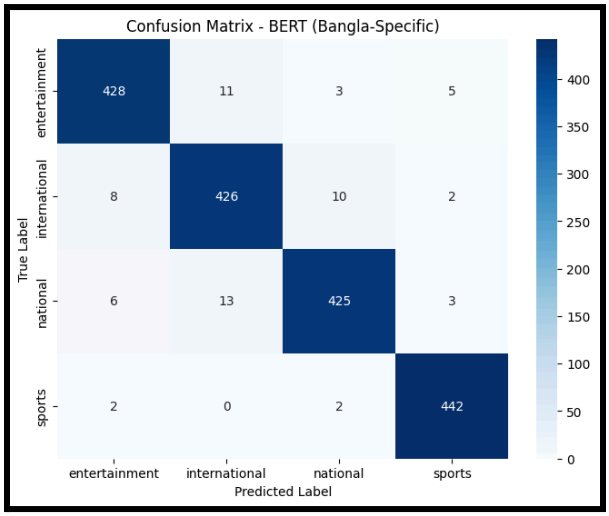


Model 6:

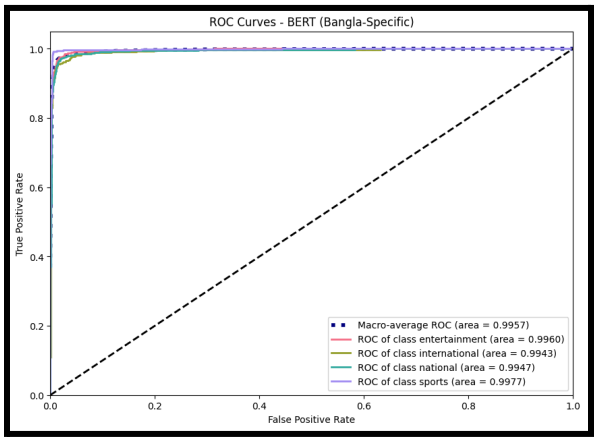
Classification Report:

```
--- BERT (Bangla-Specific) Evaluation on TEST SET ---
Test Accuracy: 0.9636

Test Classification Report:
               precision    recall  f1-score   support

entertainment     0.9640    0.9575    0.9607       447
international     0.9467    0.9552    0.9509       446
     national     0.9659    0.9508    0.9583       447
       sports     0.9779    0.9910    0.9844       446

     accuracy                         0.9636      1786
    macro avg     0.9636    0.9636    0.9636      1786
 weighted avg     0.9636    0.9636    0.9636      1786
```

Confusion Matrix:



ROC-Curves:



Model's Predictions:

```
--- SAMPLE 1 (Original Test Index: 9883) ---
RAW TEXT: বাজেটে আইএমএফের প্রেসক্রিপশন ফলো করা হয়নি: কাদের [SEP] ফাইল ছবি ২০২৪-২৫ অর্থবছরের জন্য প্রস্তাবিত বাজেটকে বাস্তবসম্মত ও গণমুখী বললেন আওয়ামী লীগ সাধারণ সম্পাদক ওবায়দুল কাদের। তিনি জানালেন, বাজেটে আন্তর্জাতিক মুদ্রা তহবিলের (আইএমএফ) প্রেসক্রিপশন ফলো করা হয়নি। জাতীয় সংসদে ২০২৪-২০২৫ অর্থবছরের বাজেট উপস...
TRUE CATEGORY: national
------------------------------
MODEL PREDICTIONS:
  Logistic Regression: national
  Random Forest:       national
  LSTM:                national
  CNN:                 national
  BERT (Multilingual): national
  BERT (Bangla-Spec):  national
============================================================
```

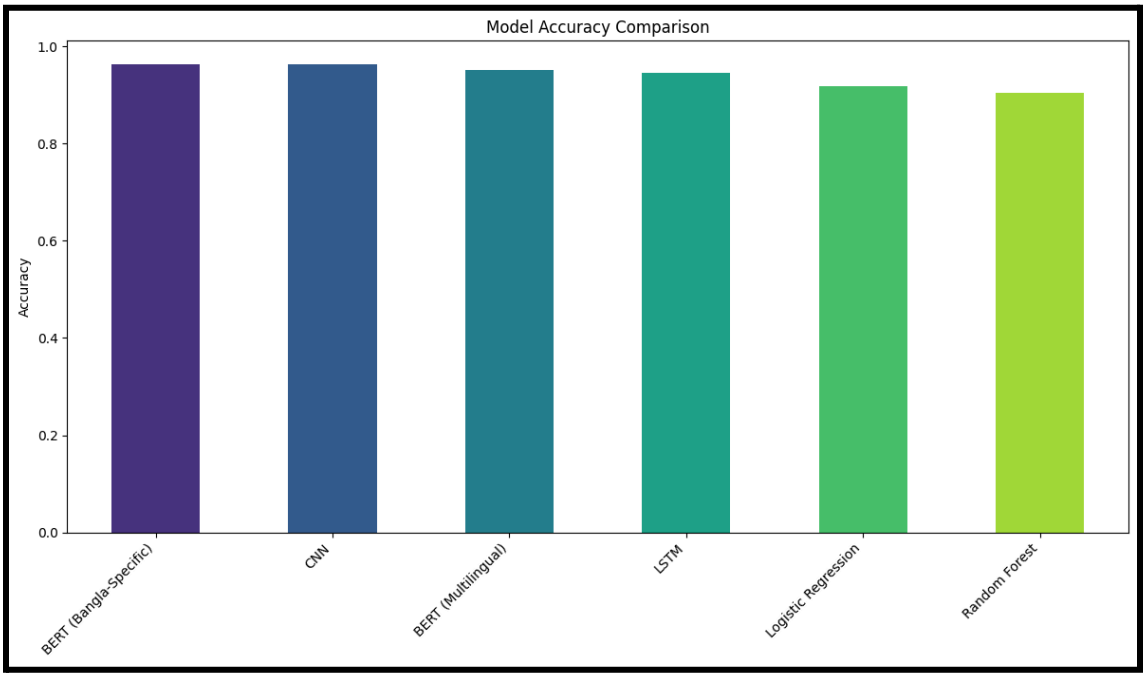
```
--- SAMPLE 2 (Original Test Index: 9607) ---
RAW TEXT: ছাগলকাণ্ড: রাজস্ব কর্মকর্তা মতিউরের সম্পদ অনুসন্ধানে দুদকের কমিটি [SEP] ফাইল ছবি ছাগলকাণ্ডে আলোচিত জাতীয় রাজস্ব বোর্ডের সদস্য মতিউর রহমানের সম্পদ অনুসন্ধানে তিন সদস্যবিশিষ্ট তদন্ত কমিটি গঠন করেছে দুর্নীতি দমন কমিশন (দুদক)। কমিটির বিষয়টি রোববার (২৩ জুন) নিশ্চিত করেছে দুদক। মতিউরের সাথে দুর্নীতিতে জড...
TRUE CATEGORY: national
------------------------------
MODEL PREDICTIONS:
  Logistic Regression: national
  Random Forest:       national
  LSTM:                national
  CNN:                 national
  BERT (Multilingual): national
  BERT (Bangla-Spec):  national
============================================================
```

Model Performance Comparison:

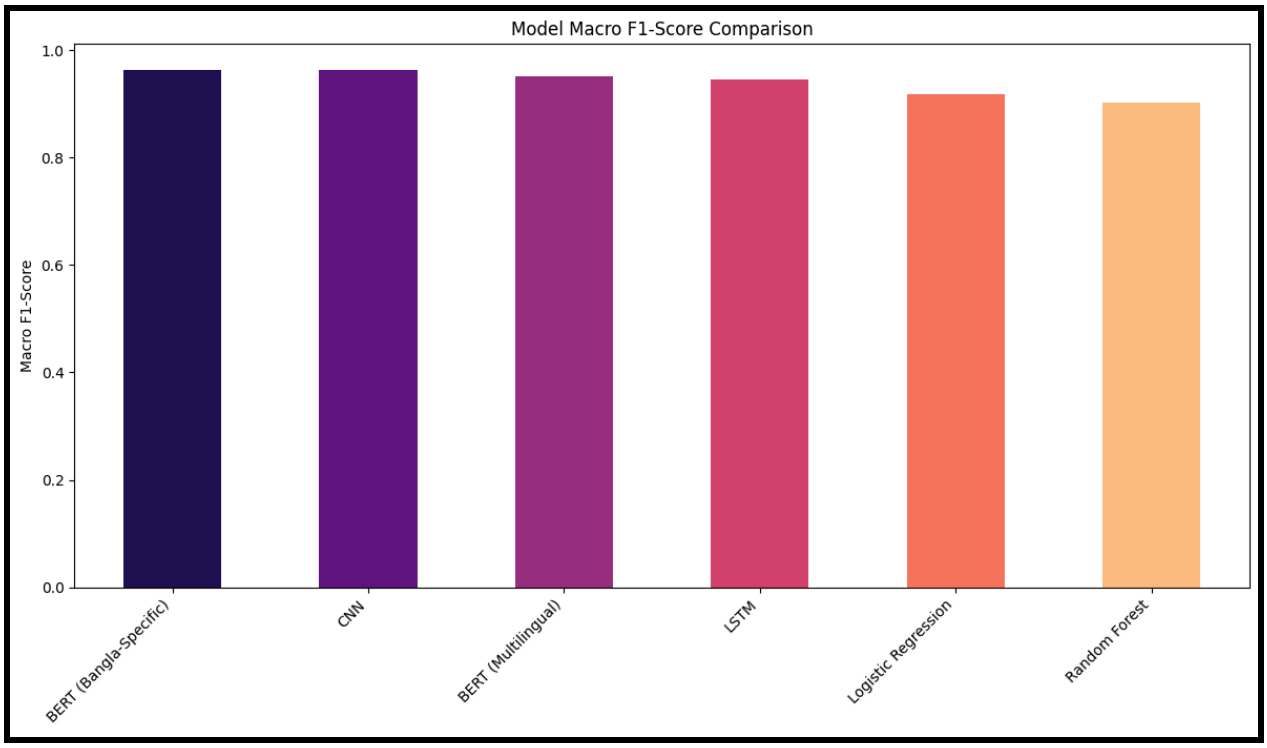
Model Performance Comparison on TEST SET

	Accuracy	Macro F1	Weighted F1	Macro AUC
BERT (Bangla-Specific)	0.963606	0.963577	0.963572	0.995656
CNN	0.963046	0.963070	0.963065	0.995506
BERT (Multilingual)	0.951848	0.951833	0.951828	0.994641
LSTM	0.945129	0.945161	0.945154	0.988968
Logistic Regression	0.917693	0.917700	0.917695	0.988579
Random Forest	0.903695	0.903463	0.903457	0.980202

Model's Accuracy:



Model's Macro F1-score:



Comparison Test: