Department of Computer Science and Engineering
Jahangirnagar University

Punctuation Prediction and Restoration for Bangla Text Using Transformer-based Learning

Presented by

Asraf Ullah Rahat
Exam Roll: 220501
Session: 2021-2022

Supervised by

Dr. Md. Ezharul Islam
Professor
Department of CSE

Tuesday, 17 September,

Slide Outline

- Introduction
- Motivation and Objective
- Literature Review
- Contribution
- Methodology
- Experimental Results
- Conclusion & Future Scope

Introduction

Bangla, the mother language of Bangladesh, is spoken by over 230 million people across Bangladesh and the Indian states of West Bengal, Tripura, and Assam. Having 300 million total speakers, Bangla is the sixth-largest language worldwide.

Punctuation marks such as commas, periods, question marks, and semicolons help to organize and segment text, guiding readers through the intended flow of information.

Missing or incorrect punctuation makes text harder to read and understand and sometimes can alter intended meaning.

Introduction(cont.)

Without Punctuations	With Punctuations	Impact
আমি বই পড়েছি আমি গল্প শুনেছি (I read a book I heard a story)	আমি বই পড়েছি। আমি গল্প শুনেছি। (I read a book. I heard a story.)	Disrupt the flow of information and create unclear sentences.
তুমি কি জানো যে তুমি আমাকে সাহায্য করতে পারবে (Do you know that you can help me)	তুমি কি জানো যে, তুমি আমাকে সাহায্য করতে পারবে? (Do you know that you can help me?)	Difficult to distinguish between a statement and a question.
আপনি কেমন আছেন আমি ভালো আছি (How are you I am fine)	আপনি কেমন আছেন? আমি ভালো আছি। (How are you? I am fine.)	Alter the perceived tone and intention of the sentence.

Motivation and Objectives

- While high-resource languages like English have seen significant advancements in educational technology, the Bangla (low-resource language) education system lags in this regard. Effective automated tools to support Bangla medium students.
- We need to develop an efficient AI for the Bangla language. Current AI tools can't support the Bangla language properly because:
  i. Lack of extensive, high-quality datasets.
  ii. Bangla has a range of dialects and regional variations. But comparatively very little research is available on the development of AI tools for Bangla.
  iii. Very complex scripts and grammar.

Motivation and Objectives(Cont.)

Original Text	Result of existing AI(ChatGPT)
তুমি জীবনকে সার্থক সুন্দর করতে চাও? ভাল কথা, কিন্তু সেজন্য তোমাকে প্রাণান্ত পরিশ্রম করতে হবে। সব তুচ্ছ করে যদি তুমি লক্ষ্যের দিকে ক্রমাগত অগ্রসর হতে পার, তবে তোমার জীবন সুন্দর হবে। আরও আছে।	তুমি জীবনকে সার্থক ও সুন্দর করতে চাও, ভাল কথা। কিন্তু সেজন্য তোমাকে প্রাণান্ত পরিশ্রম করতে হবে। সব তুচ্ছ বিষয় ভুলে যদি তুমি লক্ষ্যের দিকে ক্রমাগত অগ্রসর হতে পার, তবে তোমার জীবন সুন্দর হবে। আরও আছে,
কাজটাকে একটা তুচ্ছ, তবে বিরক্তিকর ভুল বলেই মনে হলো আমার। হলের দিকে খোলা নিচতলার পাঁচটা দরজা থেকেই নয় শুধু, বরং সিঁড়ি থেকে এবং দোতলার ব্যালকনি থেকেও পরিষ্কার দেখা যায় এমন জায়গায় রয়েছে ডাস্টপ্যানটা।	কাজটাকে একটা তুচ্ছ, তবে বিরক্তিকর ভুল বলেই মনে হলো। আমার হলের দিকে খোলা নিচতলার পাঁচটা দরজা থেকেই নয়, শুধু বরং সিঁড়ি থেকে এবং দোতলার ব্যালকনি থেকেও পরিষ্কার দেখা যায় এমন জায়গায় রয়েছে ডাস্টপ্যানটা।
কিসে হয় মর্যাদা ?দামি কাপড়,গাড়ি-ঘোড়া,না ঠাকুর–দাদার কালের উপাধিতে? না,মর্যাদা এসব জিনিসে নেই। আমি দেখতে চাই তোমার ভিতর,তোমার বাহির, তোমার অন্তর।আমি জানতে চাই, তুমি চরিত্রবান কিনা, তুমি সত্যের উপাসক কিনা।	কিসে হয় মর্যাদা? দামি কাপড়, গাড়ি, ঘোড়া, না ঠাকুর দাদার কালের উপাধিতে? মর্যাদা এসব জিনিসে নেই। আমি দেখতে চাই তোমার ভিতর, তোমার বাহির, তোমার অন্তর। আমি জানতে চাই, তুমি চরিত্রবান কিনা; তুমি সত্যের উপাসক কিনা?

Motivation and Objectives(Cont.)

- There is no research on the prediction or restoration of punctuation in Bangla language. So developing an automated systems for punctuation prediction in Bangla is therefore essential to address these gaps and take the Bangla language a step further.
- In the realm of modern natural language processing (NLP) applications, auto-mated punctuation prediction has emerged as a critical tool that will contribute to improving the level of Bangla language-based education and AI tools for Bangla.
- By improving punctuation prediction, this research will bridge the gap between machine-generated text and human-written content, ensuring that automatically produced information meets the same standards of readability and clarity as that created by humans.

Literature Review

Author(s)	Focus	Approach/Model	Key Results
Guhr et al. [1]	Multilingual Deep Learning for Punctuation	Multilingual model for sentence end and punctuation prediction	Average F1-score of 0.78 for punctuation prediction on English, German, French, and Italian texts
Tilk et al. [2]	Punctuation Restoration in Estonian Voice Transcripts	Two-stage LSTM-based model using textual and pause length information	Reduced errors by 8.8% on ASR output, 16.9% on reference text
Li et al. [3]	Sentiment Analysis	BERT-based architecture with linear classification layer	Outperformed state-of-the-art results

Literature Review(Cont.)

Author(s)	Focus	Approach/Model	Key Results
Makhija et al. [4]	Punctuation Prediction as Sequence Tagging	BERT with hybrid Bi-LSTM-CRF layer	Overall F1 score of 81.4% on the IWSLT dataset
Fang et al. [5]	Punctuation Prediction for Chinese Text	BERT, BLSTM, BERT-BLSTM-CRF	BERT-BLSTM and BERT-BLSTM-CRF outperformed all baselines
Ueffing et al. [6]	Automatic punctuation prediction in spoken and written text	CRF-based model	Relative F-score improvements of up to 26% over the baseline hidden-event language model

Contribution

- To our knowledge, there is no existing research on punctuation prediction and restoration for the Bangla language. Our research aims to fill this gap and open new avenues in this field.
- It will be an assistant of ASR for Bangla language.
- We have brought a little change in model to feed punctuation marks as a valid token.
- We present a class-wise accuracy comparison among frequently used and rarely used punctuation marks and we have tried to improve the accuracy of rarely used punctuation marks.

Methodology

We have used the BERT [7] model which is a transformer-based model.

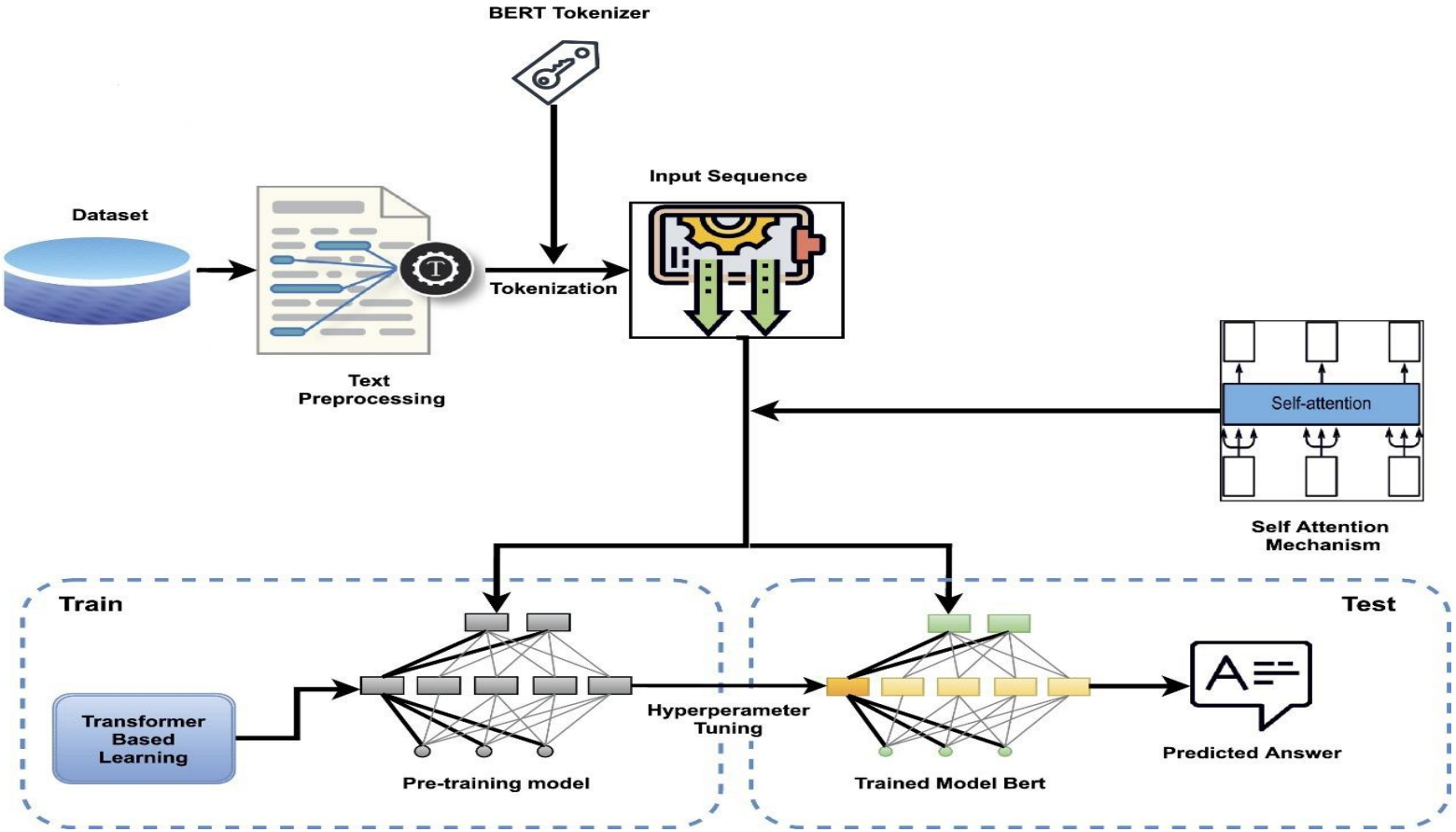

Fig: Proposed Framework

Methodology(Cont.)

- **Data Collection**

  We have collected our data from Wikipedia, NCTB textbooks, newspapers, and different online sources.

Texts
আজ আফগানিস্তানে যখন সামরিক আইন চলছে, তখন বাংলাদেশ-ভারত সিরিজ নিয়ে কথা হচ্ছে। টুইটার এবং অন্যান্য সোশ্যাল মিডিয়ায় ভারতীয় ক্রিকেটারদের নিয়ে আলোচনা চলছে। এই পরিস্থিতিতে কিছু প্রশ্ন উঠেছে যে, ভারতের ক্রিকেটাররা কি এই সিরিজের অংশ হবেন?
কাজটাকে একটা তুচ্ছ তবে বিরক্তিকর ভুল বলেই মনে হলো আমার। হলের দিকে খোলা নিচতলার পাঁচটা দরজা থেকেই নয় শুধু, বরং সিঁড়ি থেকে এবং দোতলার ব্যালকনি থেকেও পরিষ্কার দেখা যায় এমন জায়গায় রয়েছে ডাস্টপ্যানটা। জিনিসটা আসলে কী ইঙ্গিত করছে সেটা বোঝার আগে হল পার হয়ে বিরক্তিকর ডাস্টপ্যানটা সরিয়ে ফেললাম। আমার স্মরণে এলো, আধাঘন্টার মতো আগে আমার বাবা প্রবেশপথের হল পরিষ্কার করেছেন। প্রথমে এ রকম একটা ভুল আমার বাবার নামে দেখতে কষ্ট হলো আমার। তারপর নিজেকে বোঝালাম, এ রকম তুচ্ছ ভুল যখন-তখন হতে পারে। তারপর মিস কেন্টনের প্রতি আমার বিরক্তি তৈরি হলো :এ রকম একটা তুচ্ছ ভুলকে কেন্দ্র করে এভাবে অনাকাঙ্ক্ষিত বাড়াবাড়ি করে ফেললেন তিনি!

Methodology(Cont.)

- **Data Pre-processing**

Input Texts	Output Texts	Full Stop Count	Comma Count	Question Marks Count	Exclamation Marks Count
তুমি জীবনকে সার্থক সুন্দর করতে চাও ভাল কথা কিন্তু সেজন্য তোমাকে প্রাণান্ত পরিশ্রম করতে হবে সব তুচ্ছ করে যদি তুমি লক্ষ্যের দিকে ক্রমাগত অগ্রসর হতে পার তবে তোমার জীবন সুন্দর হবে আরও আছে	তুমি জীবনকে সার্থক সুন্দর করতে চাও? ভাল কথা, কিন্তু সেজন্য তোমাকে প্রাণান্ত পরিশ্রম করতে হবে। সব তুচ্ছ করে যদি তুমি লক্ষ্যের দিকে ক্রমাগত অগ্রসর হতে পার, তবে তোমার জীবন সুন্দর হবে। আরও আছে।	3	2	1	0

Methodology(Cont.)

Input Texts	Output Texts	Full Stop Count	Comma Count	Question Marks Count	Exclamati-on Marks Count
আজ আমাদের স্কুলের বার্ষিক খেলা ছিল সবাই মিলে খেলায় অংশগ্রহণ করেছিল আমাদের দলের খেলা ছিল চমৎকার। আর আজকের খেলার দিনটি খুবই আনন্দময় ছিল খেলোয়াড়রা যখন গোল করল তখন সবাই একসাথে চিৎকার করে উঠলো	আজ আমাদের স্কুলের বার্ষিক খেলা ছিল! সবাই মিলে খেলায় অংশগ্রহণ করেছিল। আমাদের দলের খেলা ছিল চমৎকার। আর, আজকের খেলার দিনটি খুবই আনন্দময় ছিল। খেলোয়াড়রা যখন গোল করল, তখন সবাই একসাথে চিৎকার করে উঠলো।	4	2	0	1

Methodology(Cont.)

- **Label Creation**

  To identify punctuation, each character in the text must be labeled. Commas, periods, question marks, and other common punctuation marks are assigned numerical labels according to a predefined mapping, while non-punctuation characters are given a default label.

Punctuations	Label
No punctuation	0
Comma(,)	1
Period (|)	2
Question Mark(?)	3
Exclamation Mark(!)	4

Methodology(Cont.)

- **Model Training**

Hyper-parameters	Value
Number of epochs	5
Batch size	32
Learning rate	2e-5

Experimental Results

- Here, input is text without punctuation and the output will be with appropriate punctuation marks.

Input	Output
বন্দরের তথ্য অনুযায়ী চলিত বছেরর ২১ আগস্ট পর্যন্ত বন্দরের জেটি ও বহির্নোঙরে অবস্থানরত ছয়িট জাহাজে চুরি-ডাকাতির ঘটনা ঘটেছ	বন্দরের তথ্য অনুযায়ী, চলিত বছেরর ২১ আগস্ট পর্যন্ত বন্দরের জেটি ও বহির্নোঙরে অবস্থানরত ছয়িট জাহাজে চুরি-ডাকাতির ঘটনা ঘটেছ।
কিসে হয় মর্যাদা দামি কাপড় গাড়ি ঘোড়া না ঠাকুর দাদার কালের উপাধিতে না মর্যাদা এসব জিনিসে নেই আমি দেখতে চাই তোমার ভিতর তোমার বাহির তোমার অন্তর।আমি জানতে চাই তুমি চরিত্রবান কিনা তুমি সত্যের উপাসক কিনা	কিসে হয় মর্যাদা ?দামি গাড়ি-ঘোড়া,না ঠাকুরদাদার কালের উপাধিতে? না,মর্যাদা এসব জিনিসে নেই।আমি দেখতে চাই তোমার ভিতর,তোমার বাহির,তোমার অন্তর।আমি জানতে চাই, তুমি চরিত্রবান কিনা,তুমি সত্যের উপাসক কিনা।

Experimental Results(Cont.)

- Our model's performance was observed under various hyperparameter settings.

Learning Rate	Batch Size	Epochs	Accuracy (%)
0.001	8	10	88.45
0.001	16	10	89.23
0.0005	8	15	90.10
0.0005	16	15	90.75
0.0001	8	20	91.22
0.0001	**16**	**20**	**91.88**

Experimental Results(Cont.)

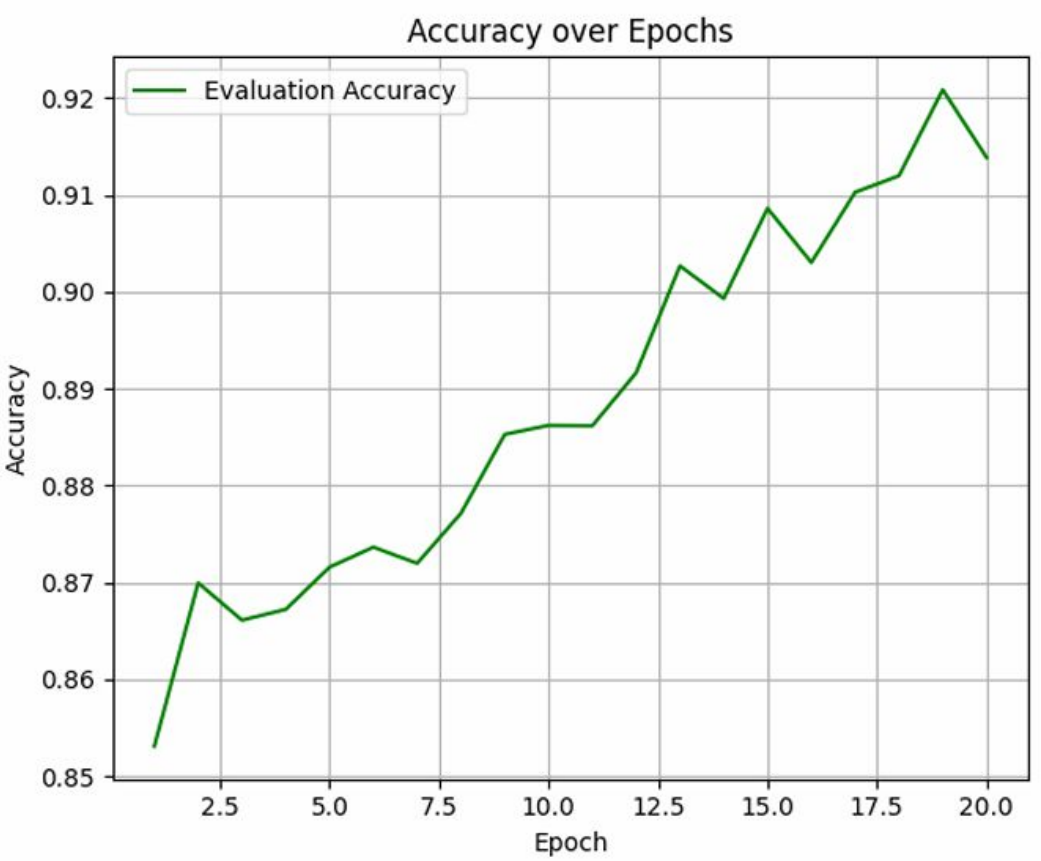

Fig: Testing accuracy over epochs

Experimental Results(Cont.)

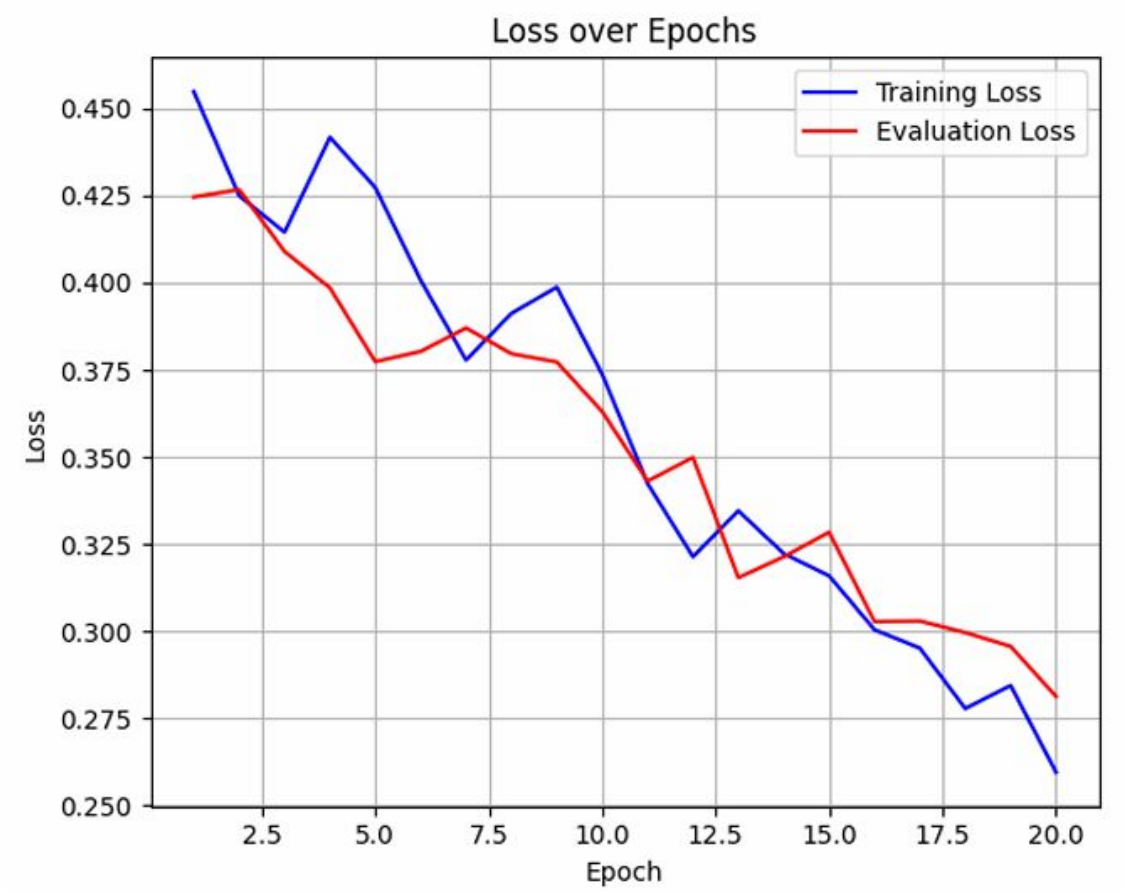

Fig: Training and testing loss over epochs

Experimental Results(Cont.)

Class-wise accuracy

We have evaluated different class accuracy using this BERT model.

Class Label	Accuracy
Comma (,)	90%
Period (।)	95%
Question Mark (?)	83%
Exclamation Mark (!)	78%

Experimental Results(Cont.)

Class-wise accuracy

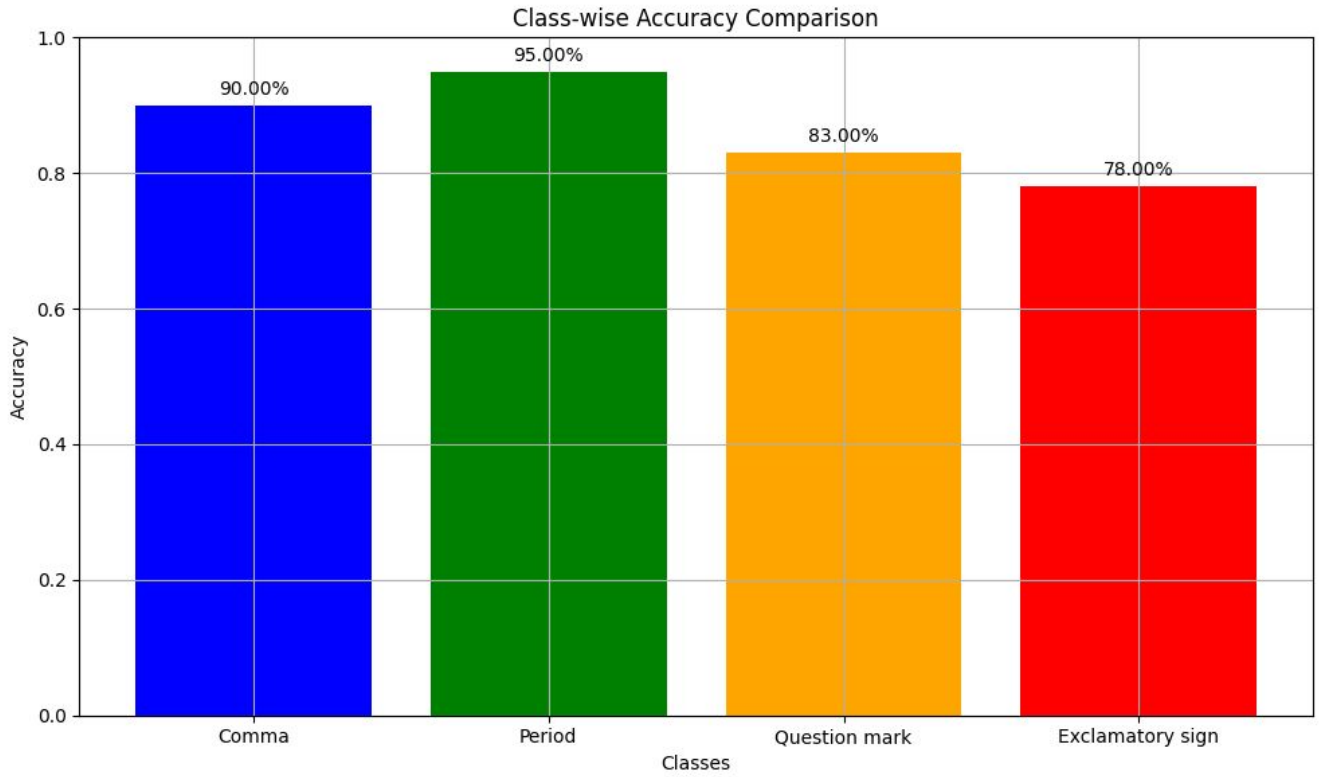

Fig: Accuracy for individual class

Experimental Results(Cont.)

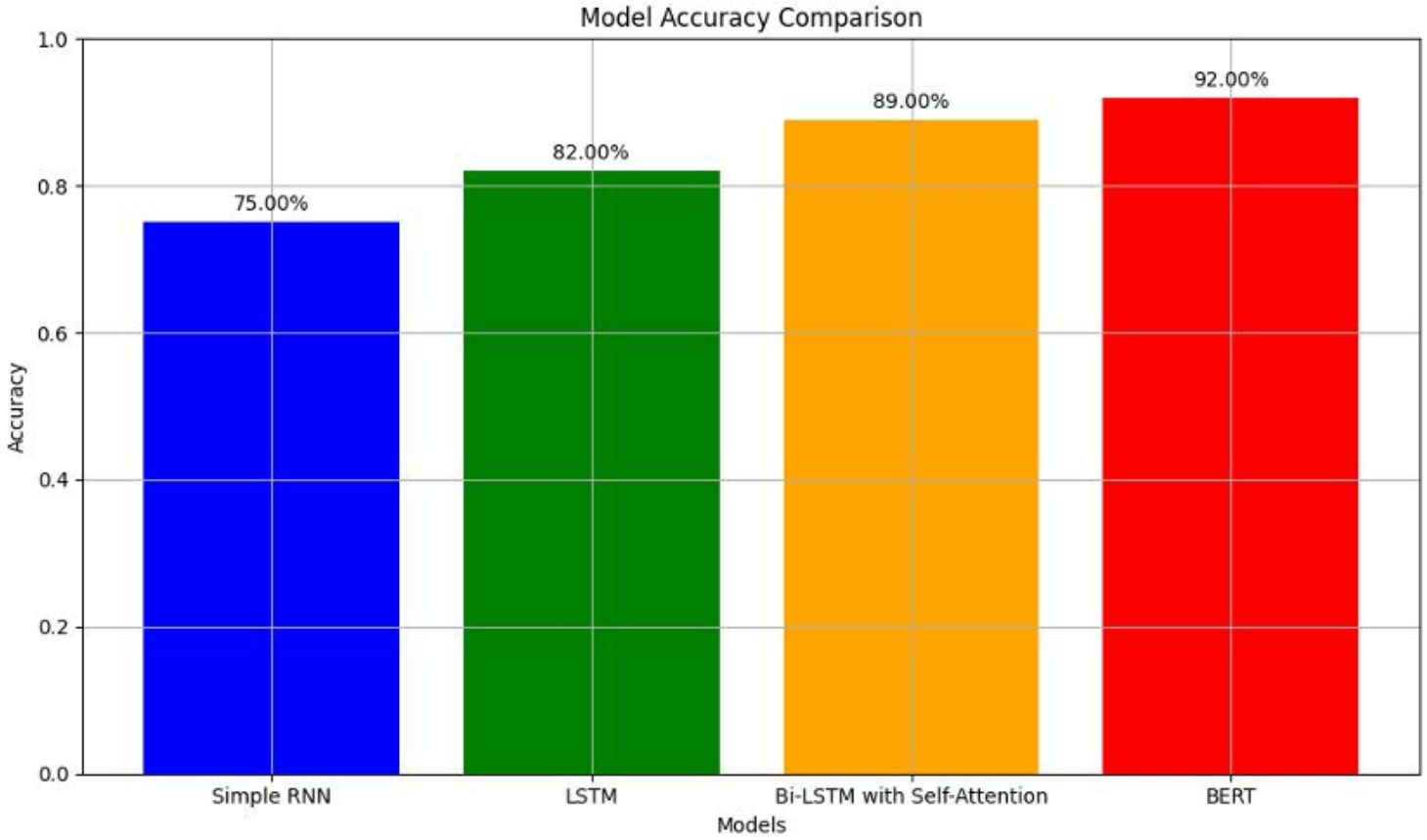

Fig: Comparison of Accuracy Metrics among NLP Models

Conclusion & Future scope

Conclusion

Given the low-resource nature of the Bangla language, our research addresses a critical gap, positioning Bangla more prominently in the global landscape of natural language processing (NLP). By leveraging a transformer-based learning approach, specifically BERT, we have demonstrated superior performance with an accuracy of 91.88%. Moreover, our work lays the groundwork for future developments in AI-driven tools for Bangla.

Conclusion & Future scope(Cont.)

Limitations

1. Lack of Diverse dataset
2. Low accuracy of rarely used punctuation marks
3. Not considering all the punctuation marks.

Conclusion & Future scope(Cont.)

Future Scope

- We aim to **develop an embedded system** based on this research to create more efficient AI applications for Bangla.
- We will consider other punctuation marks next time and enrich our dataset for a better performance.
- We will be more focused on the accuracy of **less frequently used punctuation marks.**

References

1. O. Guhr, A.-K. Schumann, F. Bahrmann, and H.-J. Böhme, "Fullstop: Multilingual deep models for punctuation prediction," in Swiss Text Analytics Conference, 2021. [Online]. Available: https://api.semanticscholar. org/CorpusID:238232903
2. O. Tilk and T. Alumäe, "Lstm for punctuation restoration in speech transcripts," in Interspeech, 2015. [Online]. Available: https://api. semanticscholar.org/CorpusID:24713327
3. X. Li, L. Bing, W. Zhang, and W. Lam, "Exploiting bert for end-to-end aspect-based sentiment analysis," arXiv preprint arXiv:1910.00883, 2019.
4. K. Makhija, T.-N. Ho, and E.-S. Chng, "Transfer learning for punctuation prediction," in 2019 Asia-Pacific Signal and Information Processing Associ ation Annual Summit and Conference (APSIPA ASC). IEEE, 2019, pp. 268–273.

References(Cont.)

5. M. Fang, H. Zhao, X. Song, X. Wang, and S. Huang, “Using bidirectional lstm with bert for chinese punctuation prediction,” in 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP). IEEE, 2019, pp. 1–5.
6. N. Ueffing, M. Bisani, and P. Vozila, “Improved models for automatic punctuation prediction for spoken and written text,” in Interspeech, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:27074188
7. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017

Thank You