Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 5, pp. 1033–1043
DOI : 10.47974/JSMS-1309

Managing token limitations with RoBERTa-large for enhanced plagiarism detection

Md. Sohail *
Principal Architect
LTIMindtree Limited
Pune
Maharashtra
India

and

Department of Computer Engineering
Savitribai Phule Pune University
Pune 411052
Maharashtra
India

Kalpana Thakre †
Department of Computer Engineering
Marathwada Mitra Mandal's College of Engineering
Savitribai Phule Pune University
Pune 411052
Maharashtra
India

Abstract

Plagiarism is a pervasive challenge in academic and digital landscapes, necessitating effective detection mechanisms. This paper presents an innovative approach utilizing advanced NLP models, specifically RoBERTa and RoBERTa-large, to enhance plagiarism detection capabilities. The seven-step methodology addresses data processing, model training, evaluation, storage, and detection, with a focus on the unique attributes of the RoBERTa-large model. Additionally, the proposed solution adeptly manages token

* *E-mail:* **Sohailmd2007@gmail.com (Corresponding Author)**
† *E-mail:* **kalpanathakre@mmcoe.edu.in**

limitations in RoBERTa-large, providing an efficient strategy for handling extended texts in plagiarism detection tasks.

This paper introduces a solution to overcome token limitations in the RoBERTa-large model, enhancing plagiarism detection accuracy. The approach involves systematic text segmentation, tokenization, and result amalgamation for efficient processing of longer texts. It tackles RoBERTa-large's constraints, ensuring optimal performance in plagiarism detection tasks.

Subject Classification: *(2010) 68M12.*

Keywords: *Plagiarism detection, NLP (Natural language processing), LLM (Large language processing), Roberta, Machine learning algorithm.*

1. Introduction

A Plagiarism poses a significant threat to the integrity of research and education in academic and digital landscapes. Copying not only undermines originality but also obstructs the progress of knowledge in an era of digital content proliferation and easy information dissemination [1-3].

This conference paper introduces an AI/ML-based solution leveraging advanced NLP models, RoBERTa and RoBERTa-large [4][5][7]. The central objective is to enhance plagiarism detection capabilities by addressing RoBERTa-large's inherent limitations, especially its token constraints with large text segments. The approach aims to fortify the educational process, promote original research, and advance the education system. The modification of RoBERTa-large is a focal point, acknowledging challenges posed by the model's maximum token limit of 512 [10].

2. Methodology

To overcome RoBERTa-large's token limitations, we employ strategic preprocessing and segmentation. Texts are systematically broken down into smaller segments using sentence tokenization [13], each processed within the 512-token limit using the RoBERTa tokenizer [4][5]. Inferences for each segment are then combined to yield conclusive results for the entire text [13]. In response to RoBERTa-large's token limitations, the proposed methodology involves a multi-step process:

2.1 *Managing Token Limitations with RoBERTa-Large*

Plagiarism undermines research and education integrity, hindering knowledge advancement. This paper presents an AI/ML-based solution

utilizing RoBERTa and RoBERTa-large models [4][5][7], aiming to improve plagiarism detection by overcoming RoBERTa-large's token constraints. The focus is on modifying RoBERTa-large to address its 512-token limit challenge [10], with the goal of enhancing education, promoting original research, and fortifying the educational system. This methodical process for handling extended texts encompasses the following steps:

A. Segmentation

Commence by partitioning the extensive text into smaller, more manageable segments or sentences. The utilization of sentence tokenization techniques or alternative methodologies aids in disassembling the text into digestible fragments.

B. Tokenization and Processing

Each segmented text is independently tokenized and processed using the RoBERTa tokenizer, ensuring compliance with the 512-token limit. This step involves careful consideration of the unique characteristics of each segment.

C. Model Inference

Processed segments are introduced into the RoBERTa model individually, generating outputs for each segment. This allows for a comprehensive understanding of the content at a granular level [12].

D. Combine Results

Merge the outputs of the model corresponding to all segments in a coherent manner to derive the conclusive result for the entire extended text. This amalgamation process may involve aggregation or summarization of the results at the segment level.

E. Model Selection and Limitation Handling

RoBERTa-large is selected for label classification support, aligning with plagiarism detection requirements [5]. To address the token limit, data is split into two sequences-Original_text and suspicious_text, each utilizing 128 tokens, totaling 512 tokens [4][5][13].

F. Additional Handling of Split Sequences

Acknowledging RoBERTa-large's limitations in processing large documents, the code introduces logic for handling split sequences [5].

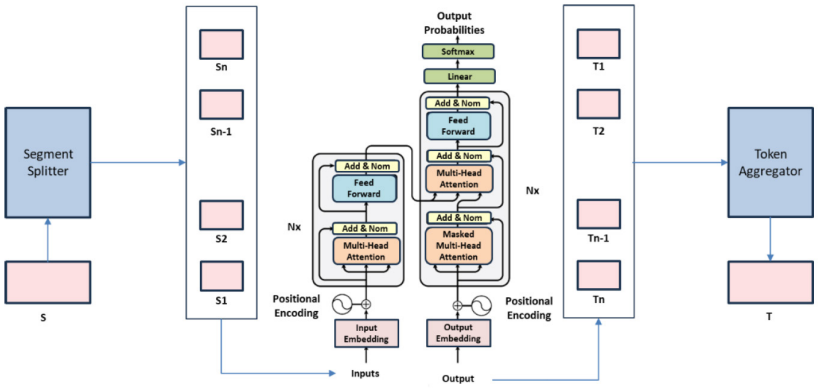

Figure 1

Transformer – model with segmentation.

Data is divided into Original_text and suspicious_text, each with a length of 128 tokens, allowing for more granular analysis of textual content [1][6] [10][13]. Transformer model is shown in Figure 1.

G. Tokenization and Dataset Preparation

The Hugging Face Transformers library facilitates data tokenization effectively [6], while the DataReader component seamlessly retrieves tokenized data from CSV files [14]. Pseudo code demonstrates sequence splitting and dataset preparation, ensuring efficient implementation of the proposed approach [4][5][6]. Additionally, the evaluation employs the PAN plagiarism corpus 2011 (PAN-PC-11), containing 11,093 source documents and 4,753 suspicious documents, providing a robust benchmark for testing automatic plagiarism detection algorithms, ensuring the methods are evaluated against a diverse set of documents.

H. Architecture diagram of Roberta-Large with Segmentation

To address RoBERTa-large's token limitations with large documents, we segment the text into smaller segments (s1, s2, ..., sn), processing each segment individually through the Input Embedding and transformer layers, generating outputs (t1, t2, ..., tn) [9][11]. These outputs are concatenated (t1 + t2 + ... + tn) to represent the entire text cohesively, aiding in dataset preparation for model training [9][11]. This systematic approach maximizes RoBERTa-large's utilization while overcoming token limits, as depicted in the architecture diagram [3][7][8].

I. Pseudo code for splitting sequences and dataset preparation

```
# Define the desired sequence length
max_length = 128

# Tokenize and prepare source data
source_input_ids = tokenizer.encode(source_text, max_length=max_length, return_tensors='pt')
source_attention_mask = tokenizer.get_attention_mask(source_input_ids)

# Tokenize and prepare suspicious data
suspicious_input_ids = tokenizer.encode(suspicious_text, max_length=max_length, return_tensors='pt')
suspicious_attention_mask = tokenizer.get_attention_mask(suspicious_input_ids)

# Concatenate the sequences
concatenated_input_ids = torch.cat((source_input_ids, suspicious_input_ids), dim=1)
concatenated_attention_mask = torch.cat((source_attention_mask, suspicious_attention_mask), dim=1)

# Prepare labels
labels = torch.tensor([label])  # Replace [label] with the actual label

# Create a dataset
dataset = TensorDataset(concatenated_input_ids, concatenated_attention_mask, labels)
```

This approach adeptly addresses the token limit of RoBERTa-large by splitting sequences into two (original and suspicious) and concatenating them before dataset preparation, ensuring effective processing of longer documents without exceeding the maximum token limit.

3. Results and Discussion

3.1 *Model training Summary Report*

This section provides a comprehensive overview of the model training process, offering insights from key reports that play a pivotal role in evaluating the performance of plagiarism detection models. The reports included are the "Model Comparison Report Analysis," "Top Performers Report," and "Loss Analysis Report." Each report contributes unique perspectives on the models' capabilities, shedding light on their accuracy, efficiency, and overall effectiveness in handling plagiarism detection tasks.

Model Comparison Report Analysis

This study compares the performance of various models, including Roberta-base, Roberta-large, Bert-base, and a modified version, Roberta-large-ms, designed to handle large document segments efficiently. The models were evaluated on three datasets: pan-pc-2009, pan-pc-2010, and pan-pc-2011. Model comparison is shown in Table I.

Table 1

Model comparison during model training

Model Comparison Report						
Sr. No	**Model**	**Dataset**	**Loss in %**	**Accuracy**	**Test Result**	**Description**
1	Roberta-base	pan-pc-2009	60.300000	65	60	Get best accuracy in 3 epochs
2	Roberta-base	pan-pc-2010	69.200000	71	70	Get best accuracy in 6 epochs
3	Roberta-base	pan-pc-2011	56.400000	72	71	Get best accuracy in 3 epochs
4	Roberta-large	pan-pc-2009	59.100000	76	72	Get best accuracy in 5 epochs
5	Roberta-large	pan-pc-2010	67.600000	72	72	Get best accuracy in 5 epochs
6	Roberta-large	pan-pc-2011	68.100000	73	69	Get best accuracy in 6epochs
7	Bert-base	pan-pc-2009	68.100000	71	65	Get best accuracy in 4 epochs
8	Bert-base	pan-pc-2010	71.300000	66	55	Get best accuracy in 5 epochs
9	Bert-base	pan-pc-2011	69.900000	69	65	Get best accuracy in 5 epochs
10	Roberta-large-ms	pan-pc-2009	56.400000	80	76	Get best accuracy in 6 epochs
11	Roberta-large-ms	pan-pc-2010	59.100000	76	80	Get best accuracy in 7 epochs
12	Roberta-large-ms	pan-pc-2011	67.600000	74	75	Get best accuracy in 7 epochs

Analysis:

Roberta-large-ms consistently achieves superior accuracy (74% to 80%) and excels in handling large document segments. Roberta-base and Bert-base demonstrate competitive performance with variability across datasets, while Roberta-large shows improvement over the base model but is surpassed by the modified version in accuracy and stability, underscoring the significance of the modifications made.

Top Performers Report

This report focuses on the top-performing models across various datasets, emphasizing their loss, accuracy, and overall test results. Top performer model result shown in Table 2.

Table 2

Top performer model based on model training results

Top Performers Report						
Model	**Sr. No**	**Dataset**	**Loss in %**	**Accuracy**	**Test Result**	**Description**
Bert-base	7	pan-pc-2009	68.100000	71	65	Get best accuracy in 4 epochs
Roberta-base	3	pan-pc-2011	56.400000	72	71	Get best accuracy in 3 epochs
Roberta-large	4	pan-pc-2009	59.100000	76	72	Get best accuracy in 5 epochs
Roberta-large-ms	10	pan-pc-2009	56.400000	80	76	Get best accuracy in 6 epochs

Analysis:

Roberta-large-ms emerges as the top performer, consistently achieving the highest accuracy (80%) and demonstrating robustness in handling large document segments. Roberta-base showcases competitive accuracy (72%) and stability, excelling particularly in pan-pc-2011. Roberta-large maintains strong performance with an accuracy range of 72% to 76%, performing well in pan-pc-2009.

Loss Analysis Report:

This report provides insights into the mean, minimum, and maximum loss values for each model, offering a comprehensive view of their convergence and stability. Loss analysis shown in Table 3.

Table 3

Loss analysis during model training

Loss Analysis Report			
Model	**mean**	**min**	**max**
Bert-base	69.766667	68.100000	71.300000
Roberta-base	61.966667	56.400000	69.200000
Roberta-large	64.933333	59.100000	68.100000
Roberta-large-ms	61.033333	56.400000	67.600000

Analysis:

Bert-base exhibits high mean loss (69.77%) and variability in convergence (68.1% to 71.3%). Roberta-base shows lower mean loss (61.97%) with consistent but variable convergence (56.4% to 69.2%). Model

Table 4

Model probability score against the suspicious document

Models With Probability For Each Suspicious Document					
Sr. No.	**Suspicious Document/Model**	**Bert-base**	**Roberta-base**	**Roberta-large**	**Roberta-large-ms**
1	suspicious-document00003.txt	0.320000	0.430000	0.450000	0.470000
2	suspicious-document20001.txt	0.410000	0.670000	0.320000	0.670000
3	suspicious-document00011.txt	0.650000	0.550000	0.580000	0.620000
4	suspicious-document01000.txt	0.320000	0.560000	0.600000	0.620000
5	suspicious-document03000.txt	0.430000	0.530000	0.340000	0.580000
6	suspicious-document02000.txt	0.440000	0.670000	0.570000	0.650000
7	suspicious-document00200.txt	0.370000	0.350000	0.670000	0.680000
8	suspicious-document00100.txt	0.450000	0.450000	0.500000	0.570000
9	suspicious-document0010.txt	0.310000	0.570000	0.400000	0.540000
10	suspicious-document00500.txt	0.400000	0.580000	0.670000	0.560000

probability score shown in Table 4. Roberta-large presents moderate mean loss (64.93%) and moderate variability (59.1% to 68.1%), while Roberta-large-ms displays the lowest mean loss (61.03%) and enhanced stability with a narrower range (56.4% to 67.6%), correlating with superior accuracy.

3.2 *Plagiarism Result Summary Report*

This section combines information from the previous reports, highlighting the models' performance, the best-performing model for each document, and summary statistics. The aim is to offer a consolidated overview of the plagiarism detection.

Models with probability for each Suspicious document

This section outlines the probability scores assigned by each plagiarism detection model to individual suspicious documents within the dataset. The probability values, ranging from 0 to 1, reflect the confidence of each model in identifying potential instances of plagiarism for each document.

Analysis:

The probability scores among models vary, highlighting nuanced approaches and inherent differences in their learning mechanisms, such as Bert-base assigning 0.320 and Roberta-large-ms assigning a slightly higher 0.470 for "suspicious-document00003.txt".

Summary Statistics for Each Model

This section provides summary statistics for each model, including mean, median, minimum, and maximum probability scores. These statistics offer a holistic view of the models' overall performance across all suspicious documents. Probability Statistic result shown in Table 5

Table 5

Probability Statistic for each model

Summary Statistics for Each Model

Model	Mean Probability	Median Probability	Min Probability	Max Probability
Bert-base	0.410000	0.405000	0.310000	0.650000
Roberta-base	0.536000	0.555000	0.350000	0.670000
Roberta-large	0.510000	0.535000	0.320000	0.670000
Roberta-large-ms	0.596000	0.600000	0.470000	0.680000

Analysis:

Analyzing the summary statistics allows us to grasp the overall behavior of each model. For instance, Bert-base exhibits a mean probability of 0.410, indicating a moderate level of confidence, while Roberta-large-ms demonstrates a higher mean probability of 0.596, suggesting greater overall confidence in its predictions.

4. Conclusion

This paper presents a robust methodology for managing token limitations in RoBERTa-large, specifically tailored for plagiarism detection tasks. The focus on modifying RoBERTa-large to handle larger segments of text addresses critical challenges in the field of plagiarism detection, offering a nuanced and effective solution that upholds academic integrity in the face of evolving digital landscapes. The proposed approach enables the efficient handling of longer texts, ensuring optimal performance without compromising on accuracy. Future work may explore alternative models designed explicitly for extended sequences and document-level tasks.

5. Future Work

Reflecting on my current efforts, I have effectively managed large sequences of documents by dividing them into multiple segments. Despite efforts to enhance computational power, performance has been compromised. It is imperative to acknowledge the existence of models explicitly crafted for processing longer sequences, such as the Longformer. This model is adept at efficiently handling documents with thousands of tokens.

In situations where tasks routinely involve processing extensive texts, it is prudent to explore alternatives or employ models like the Longformer tailored for document-level tasks. In my forthcoming endeavors, I intend to place greater emphasis on the design and implementation of the Longformer model within my solution. This strategic approach aims to optimize the handling of lengthy documents and elevate the overall performance of the system.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization (2016). arXiv preprint arXiv:1607.06450.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate (2014). CoRR, abs/1409.0473.

[3] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc V. Le. Massive exploration of neural machine translation architectures (2017). CoRR, abs/1703.03906.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint arXiv:1810.04805.

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. Roberta: A robustly optimized BERT approach (2019). arXiv preprint arXiv:1907.11692.

[6] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning (2017). arXiv preprint arXiv:1705.03122v2.

[7] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation (2014). CoRR, abs/1406.1078.

[8] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding (2017). arXiv preprint arXiv:1703.03130.

[9] Ofir Press and Lior Wolf. Using the output embedding to improve language models (2016). arXiv preprint arXiv:1608.05859.

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision (2015). CoRR, abs/1512.00567.

[11] Muttlak, Hassen A. "Estimation of parameters in a multiple regression model using rank set sampling." *Journal of Information and Optimization Sciences* 17.3 : 521-533 (1996).

[12] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling (2016). arXiv preprint arXiv:1602.02410.

[13] Altameem, Ayman, et al. "P-ROCK: a sustainable clustering algorithm for large categorical datasets." *Intell. Autom. Soft Comput* 35.1 : 553-566 (2023).

[14] Wazalwar, Sampada S., and Urmila Shrawankar. "Interpretation of sign language into English using NLP techniques." *Journal of Information and Optimization Sciences* 38.6 : 895-910 (2017).

Received May, 2024