Detecting Synthetic Text: Applying and Analyzing a Watermarking Algorithm on a Pre-Trained Summarization Model

Seyedehfatemeh (Fateme) Karimi

Date:

August 14, 2023

I. Title of Proposed Research Project: Detecting Synthetic Text: Applying and Analyzing a Watermarking Algorithm on a Pre-Trained Summarization Model

II. Introduction:

Large language models (LLMs) have the potential to generate text with human-like capabilities, but there are concerns about their misuse for malicious purposes such as social engineering, fake news creation, and academic cheating. To address these concerns, this research proposal aims to apply and analyze the watermarking algorithm proposed in "A Watermark for Large Language Models" on a pre-trained summarization model (T5, BART, Flan or others) and a summarization dataset (CNN, Xsum or other). The watermarking algorithm embeds signals into generated text that are invisible to humans but algorithmically detectable from a short span of tokens. By applying this algorithm to the pre-trained summarization model and summarization dataset, we can assess the effectiveness of the watermark in detecting machine-generated text and its potential applications in the field of natural language processing [1].

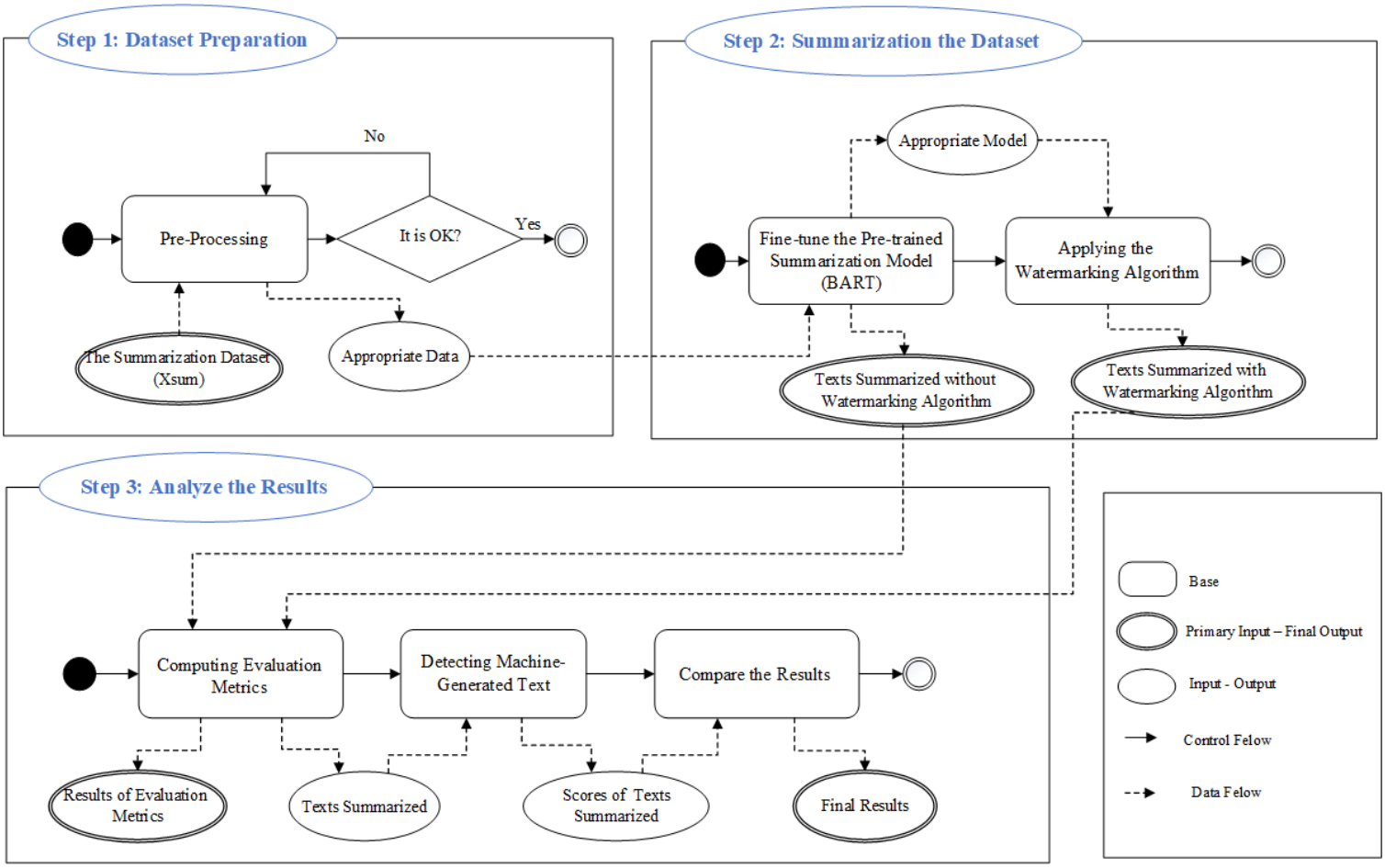

Figure 1. A Schematic Overview of the Proposed Method

III. Objectives:

The main objectives of this research proposal are as follows:

- ✓ Apply the watermarking algorithm from "A Watermark for Large Language Models" to a pre-trained summarization model (T5, BART, Flan or others) and a summarization dataset (CNN, Xsum or other).
- ✓ Generate watermarked summaries using the pre-trained summarization model and evaluate the quality of the watermarked text.
- ✓ Analyze the watermark to determine its effectiveness in detecting machine-generated text.
- ✓ Compare the results of the watermarking algorithm to the results of the same model without the watermark.

IV. Methodology:

As shown in the Figure 1, the study consists of three stages. In this section, we propose proposed method and explain its details.

- ✓ **Dataset Preparation:** Pre-process the summarization dataset to remove unnecessary information and ensure the text is in a suitable format for the pre-trained summarization model. We use the Xsum dataset to conduct this research.

- ✓ **Use of the Pre-Trained Summarization Model:** Fine-tune the pre-trained summarization model on the summarization dataset to generate summaries. Our research is based on BART summarization model. The parameters utilized are presented in Table 1.

Table 1. The parameters of the BART model for text summarization without applying the watermark algorithm

No	Parameter Name	Value
1	Tokenizer	facebook/bart-large
2	Model	facebook/bart-large
3	max_length	200
4	num_beam	4

- ✓ **Applying the Watermarking Algorithm:** Apply the watermarking algorithm proposed in "A Watermark for Large Language Models" to the generated summaries. Embed signals into the text that are invisible to humans but algorithmically detectable from a short span of tokens. The code[1] presented in the basic research is used as the basis for this study, and modifications is made to the code based on the research

[1] https://github.com/jwkirchenbauer/lm-watermarking/tree/main

objective. The BART model is utilized for document summarization, and the parameters utilized are listed in Table 2.

Table 2. The parameters of the BART model for text summarization with applying the watermark algorithm

No	Parameter Name	Value
1	Tokenizer	facebook/bart-large
2	Model	facebook/bart-large
3	max_length	200
4	num_beam	4
5	gamma	0.25
6	delta	0.2
7	seeding_scheme	simple_1
8	select_green_tokens	True

- ✓ **Comparison and Evaluation:** Compare the results of the watermarking algorithm to the results of the same model without the watermark. The results of the Rough and Bleu metrics are stated in Table 3. The results presented in this study are related to 203 documents available in the Xsum_Train dataset. The runtime for obtaining the summaries of these 203 documents in both watermarked and non-watermarked modes is more than 5 hours.

Table 3. The results of the Rough and Bleu metrics

No	Average BLEU Score	Average ROUGE Score								
		ROUGE-1			ROUGE-2			ROUGE-L		
		Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
1	0.7360463761115246	0.73	0.75	0.73	0.63	0.64	0.63	0.58	0.60	0.58

- ✓ **Watermark Analysis:** Analyze the watermark to assess its effectiveness in detecting machine-generated text. Use the watermark detection algorithm to detect the watermark in the generated summaries. The results related to the Z-score metric have been included in the dataset of results. This metric has been measured for each of the summarized documents. According to the base article, documents with a Z-score greater than 4 are considered as documents on which the watermark algorithm has been applied. However, the threshold for the Z-score can vary depending on the research objective.

V. Expected Outcomes:

- ✓ Watermarked summaries generated by the pre-trained summarization model.

- ✓ Analysis of the watermark's effectiveness in detecting machine-generated text.
- ✓ Comparison of results between watermarked and non-watermarked models.
- ✓ Statistical measure of confidence in detecting the watermark.

VI. Conclusion:

This research proposal aims to apply and analyze the watermarking algorithm proposed in "A Watermark for Large Language Models" on a pre-trained summarization model (BART) and summarization dataset (Xsum). By embedding signals into generated text that are invisible to humans but algorithmically detectable, the watermarking algorithm can potentially detect machine-generated text and mitigate potential harms of large language models. The proposed research will provide insights into the effectiveness of the watermarking algorithm in detecting machine-generated text and its potential applications in the field of natural language processing.

VII. Reference

[1] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A Watermark for Large Language Models," 2023, [Online]. Available: http://arxiv.org/abs/2301.10226