Token Prediction as Implicit Classification for Generative AI Authorship Verification

Notebook for the PAN Lab at CLEF 2024

Zhanhong Ye[1,†], Yutong Zhong[1], Zhen Huang[2] and Leilei Kong[1,†]

[1]Foshan University, Foshan, Guangdong, China
[2]South China Normal University, Guangzhou, Guangdong, China

Abstract

This paper presents a method leveraging Next Token Prediction as Implicit Classification for Voight-Kampff Generative AI Authorship Verification. The rationale behind this approach is that token prediction can effectively perform text classification tasks. Consequently, we utilize the Token Prediction method to directly identify whether the input text was authored by a specific AI model or by a human. We assessed the effectiveness of our method using the Generative AI Authorship Verification datasets provided by PAN. We then selected model weights that demonstrated the best performance on the dataset given by PAN. Finally, on the test set, our performance metrics at the Minimum, 25-th Quantile, Median, 75-th Quantile, and Max were 0.527, 0.896, 0.922, 0.926, and 0.947 respectively.

Keywords

PAN 2024, Voight-Kampff Generative AI Authorship Verification 2024, Next token prediction

1. Introduction

In recent years, generative LLMs have gained recognition for their impressive ability to produce coherent language across different domains. Consequently, detecting machine-generated text has become increasingly vital. The Generative AI Authorship Verification task regarded as detecting machine-generated text task involves two texts, one authored by a human and one by a machine. The primary objective is to determine which of the two texts was written by a human and which was generated by a machine. Furthermore, the Generative AI Authorship Verification task can aid in ensuring the authenticity of information is critical, such as legal proceedings.

Research [1] utilizes Token Prediction as an Implicit Classification for Generative AI Authorship Verification. By assigning distinct tokens to different labels and reformulating the multi-class classification task into a next-token prediction task, this method identifies whether the input sentence was generated by a particular model or authored by a human [1]. The purpose of this approach is to leverage the model's next-token prediction capability for this specific task.

Recent studies [2] have employed the fine-tune transformer-based method, which achieved the LLMs-generated text detection task by training transformer-based classifiers. However, one of the biggest challenges in fine-tuning transformer-based methods is not to directly leverage the next-token prediction capability of the model for this particular task [1]. Fine-tune transformer-based method will increase the gap between downstream tasks and pre-training tasks compared to next-token prediction [3]. Hence here are better solutions than simply fine-tuning transformer-based methods.

In this paper, we leverage research [1] to predict whether a given sample text is authored by a human or paraphrased by a machine. Unlike the fine-tuning transformer-based method, we employ the Token Prediction as an Implicit Classification approach. This involves establishing a bijection $f:Y \to \mathcal{Y}$,

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

† corresponding author

chinwang.yip@gmail.com (Z. Ye); yutongz115@gmail.com (Y. Zhong); 20222632026@m.scnu.edu.cn (Z. Huang); kongleilei@fosu.edu.cn (L. Kong)

0009-0001-4094-006X (Z. Ye); 0009-0003-1694-9800 (Y. Zhong); 0009-0000-6220-4656 (Z. Huang); 0000-0002-4636-3507 (L. Kong)

where $Y \subset \Sigma$. $\mathcal{Y}$ serves as proxy labels such as 'human', 'GPT-3.5', etc. Y represents the ground truth label. The model then predicts the corresponding proxy labels based on the input text.
We have established two sets of proxy labels which are proxy labels in method 1 and proxy labels in method 2.
In method 1, the proxy labels can be translated into three outcomes: one indicating human authorship, one indicating AI model rewrites, and one indicating undecidable.
Method 2, proxy labels are translated into two outcomes: one for human authorship and one for AI model rewrites. This differentiation allows us to determine whether the text is human-authored, machine-generated, or falls into another category.
In detail, the model comprises two parts. The first part is the long-T5 [4] model, which encodes the input text. The second part is a linear layer designed to project the output of long-T5 onto a dimension equivalent to the vocabulary size. This projects the probabilities of the proxy labels, thereby determining whether the input text under examination was generated by a model, authored by a human or undecidable.

2. Network Architecture

First, the language model is presented with a series of sentences to be tested, each consisting of tokens from E_1 to E_n and $E_{<s>}$. The goal is to utilize the longT5 model to implement the Generative AI Authorship Verification task. The core feature of the model is the method of next-token prediction. After inputting the tokens from E_1 to E_n and $E_{<s>}$ into longT5, it obtains the probabilities of the proxy labels. The predicted proxy labels for each sentence are then determined by selecting the label with the highest probability. Then, we convert the proxy labels into the final result, determining whether the text was authored by a human or paraphrased by a machine. According to the model shown in Figure 1, it comprises a longT5 backbone, a next-token prediction layer, and a filter. The first component is the longT5 backbone, which is used to encode the sentences under examination. Following the next-token prediction layer in method 2, where linear layers map the output of longT5 to a dimension equivalent to the vocabulary size, enabling the calculation of probabilities for each proxy label.
In method 2, the filter selects the probabilities corresponding to the proxy labels from the output of the next-token prediction layer, which are then processed through a softmax layer. Finally, the proxy labels with the highest probability are chosen, which is then translated into one of two outcomes: whether the text under examination was generated by a specific model or authored by a human.
Returning to method 1, it is similar to method 2 but it identifies the text by obtaining the probabilities corresponding to the proxy labels from the next-token prediction layer. In method 1, after obtaining the proxy labels, we translate them into two outcomes: one indicating human authorship and the other AI model rewrites. For method 1, in addition to these two outcomes, we include an additional result labeled as undecidable, making three possible outcomes. The detailed process is described in section 2.1.
Overall, the primary loss function $\mathcal{L}$ can be defined as follows.

$$\mathcal{L} = \mathcal{L}_{NLL} = -logP(\mathcal{Y}_i|S_i;\theta) \tag{1}$$

The loss $\mathcal{L}_{NLL}$ is negative log-likelihood to optimize the longT5 and next-token prediction layer, S_i means the sentence under examination, θ mean the whole model's parameters, and y means the ground truth labels.

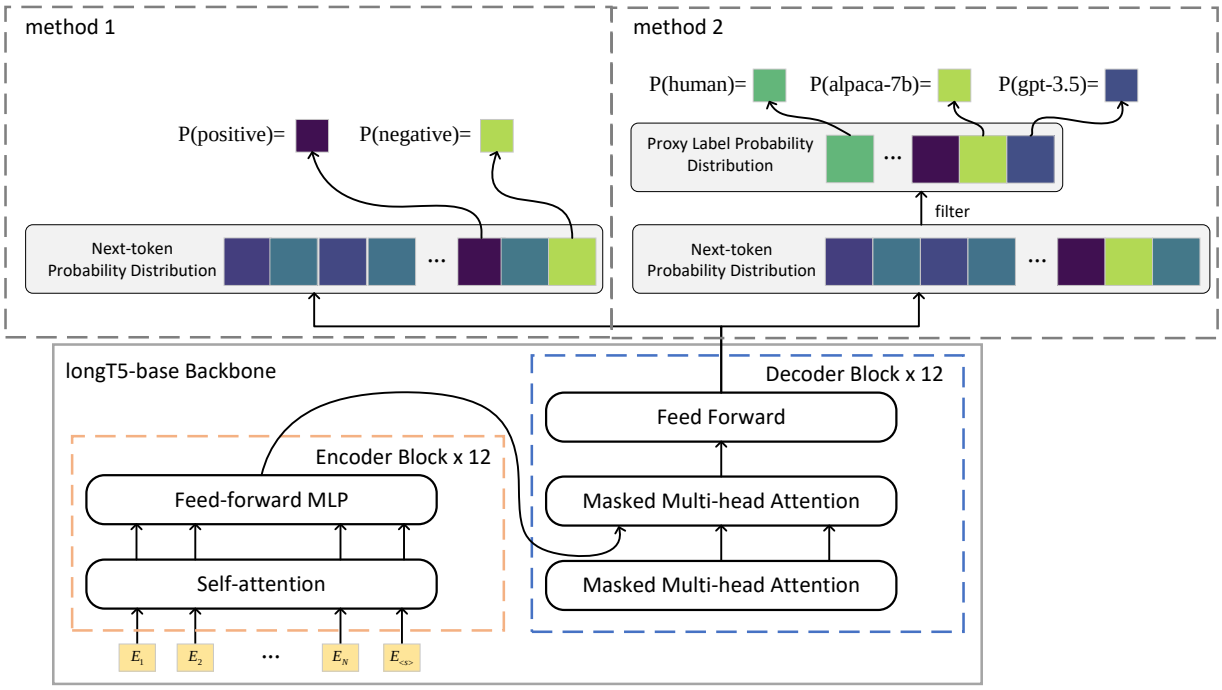

Figure 1: Figure1 Model Architecture

2.1. next-token prediction

For method 2, we assign a special token "<extra_id_0>" as the proxy label for human-authored text. For other models we designate similar tokens such as "<extra_id_1>", "<extra_id_2>", ..."<extra_id_n>", where $n \leq k$ and k represents the number of models involved in PAN dataset [5, 6].
In method 1, human-authored texts are tagged with the word "positive" as the proxy label, while all texts rewritten by AI models are labeled "negative". If the highest probability in the next token probability distribution does not fall on either "positive" or "negative", the result is deemed "undecidable".
Both methods involve the model predicting the probabilities of the proxy labels and then converting proxy labels into the actual prediction results.
Next, we measure the token length of each human-authored or model-generated text. Our statistical analysis reveals that the vast majority of text lengths are within 2048 tokens.
Firstly, the PAN organization has provided datasets for Generative AI Authorship Verification, which include multiple texts authored by humans and subsequently rewritten by various models.
Give a batch name as $\mathcal{B}$. The contents of $\mathcal{B}$ can define as $\{(S_1, \mathcal{Y}_1), (S_2, \mathcal{Y}_2)...(S_i, \mathcal{Y}_i)\} \in \mathcal{B}$, where S_i means the sentence under examination, and $\mathcal{Y}_i$ is the proxy label.
During training, we feed the $\mathcal{B}$ into the pre-training model which is composed of the transformer [7] block to get the corresponding hidden state $\mathcal{H}_i$. After obtaining the hidden state $\mathcal{H}_i$ we use the next-token prediction layer and softmax layer to obtain the probabilities for all tokens in the vocabulary. That is,

$$\varphi_i = (\mathcal{Y}_i^1, \mathcal{Y}_i^2, ...\mathcal{Y}_i^V) = (\frac{e^{(\phi(\mathcal{H}_i)^1)}}{\sum_{v=1}^{V} e^{(\phi(\mathcal{H}_i)^V)}}, \frac{e^{(\phi(\mathcal{H}_i)^2)}}{\sum_{v=1}^{V} e^{(\phi(\mathcal{H}_i)^V)}}, ..., \frac{e^{(\phi(\mathcal{H}_i)^V)}}{\sum_{v=1}^{V} e^{(\phi(\mathcal{H}_i)^V)}}) \tag{2}$$

where φ_i is the soft label of sample i, v indicates the position of a token within the vocabulary, V represents the total number of tokens in vocabulary, $\mathcal{Y}_i^V$ represents the probability of the V-th word in vocabulary and $\mathcal{Y}_i$ means proxy label. Then we calculate the negative log-likelihood loss for classification.

$$L_{nll} = -logP(\mathcal{Y}_i|\varphi_i, \theta) \tag{3}$$

In the inference phase, for method 1, after obtaining φ_i , we convert φ_i into three predictive outcomes.

$$\hat{y} = \begin{cases} 1 & arg \max\limits_{\mathcal{Y} \in \mathcal{Y}_i^V} \varphi_i = a \\ 0 & arg \max\limits_{\mathcal{Y} \in \mathcal{Y}_i^V} \varphi_i = b \\ 0.5 & otherwise \end{cases} \tag{4}$$

In method 1, a represents the position of the word “positive” in the vocabulary, while b represents the position of the word "negative". $\hat{y}$ represent predict label. $\hat{y}$ = 1 indicates text authored by humans, $\hat{y}$ = 0 indicates text rewritten by a machine, and $\hat{y}$ = 0.5 indicates ’undecidable’ when a clear determination cannot be made.
For method 2, we initially obtain the output from the next-token prediction layer.

$$\phi(\cdot) = (\phi(\mathcal{H}_i)^1, \phi(\mathcal{H}_i)^2, ..., \phi(\mathcal{H}_i)^{\mathrm{V}}) \tag{5}$$

where $\phi(\cdot)$ indicates the output of the next-token prediction layer and V is vocabulary size. We then use a filter to select the outputs associated with all the special tokens(proxy label tokens).

$$\phi(\cdot)' = (\phi(\mathcal{H}_i)^1, \phi(\mathcal{H}_i)^2, ..., \phi(\mathcal{H}_i)^k) \tag{6}$$

where $\phi(\cdot)'$ indicates the output of the filter and k represents the number of all special tokens. After passing through the softmax layer, we obtain the probability distribution of proxy label tokens.

$$\varphi_i' = (\mathcal{Y}_i^1, \mathcal{Y}_i^2, ...\mathcal{Y}_i^k) = (\frac{e^{(\phi(\mathcal{H}_i)')^1}}{\sum_{j=1}^k e^{(\phi(\mathcal{H}_i)')^j}}, \frac{e^{(\phi(\mathcal{H}_i)')^2}}{\sum_{j=1}^k e^{(\phi(\mathcal{H}_i)')^j}}, ..., \frac{e^{(\phi(\mathcal{H}_i)')^k}}{\sum_{j=1}^k e^{(\phi(\mathcal{H}_i)')^j}}) \tag{7}$$

where $j \in k$ and φ_i' represent probability distribution of proxy label tokens. Finally, we convert $\varphi_i{}'$ into two predictive outcomes:

$$\hat{y} = \begin{cases} 0 & arg \max\limits_{\mathcal{Y} \in \mathcal{Y}_i^V} \varphi_i = c \\ 1 & otherwise \end{cases} \tag{8}$$

where $c \in \{1 \dots k\}$ indicates the special tokens, $\hat{y}$ = 1 indicates text authored by humans, and 0 indicates text rewritten by a machine.

3. Experiments and Result

3.1. Experience setting

In this work, we utilize the longT5 model for classification, which consists of 12 transformer layers, with a hidden size of 768. As for the next-token prediction layer, we use randomly initialized parameters before training. For method 1, the training parameters are set with 10 epochs, a batch size of 64, and a learning rate of 5e-4. For method 2, the settings are 15 epochs, a batch size of 16, and a learning rate of 8e-4. Both method’s maximum token length is set to 2048. All experiments are conducted on an NVIDIA A800 GPU with 80GB of memory.

3.2. Results

We will conduct two experiments using token prediction as an implicit classification for both method 1 and method 2. After training with these methods, the resulting model weights from both experiments will be submitted to the TIRA platform [8] to obtain scores. Table 1 and 2 displays our test set results reported to the TIRA platform.

Table 1 shows the summarized results averaged (arithmetic mean) over 10 variants of the test dataset. Each dataset variant applies one potential technique to measure the robustness of authorship verification approaches, e.g., switching the text encoding, translating the text, switching the domain, manual obfuscation by humans, etc.

Table 2 shows the results, initially pre-filled with the official baselines provided by the PAN organizers and summary statistics of all submissions to the task (i.e., the maximum, median, minimum, and 95-th, 75-th, and 25-th percentiles over all submissions to the task).

Table 1
Overview of the accuracy in detecting if a text is written by an human in task 4 on PAN 2024 (Voight-Kampff Generative AI Authorship Verification). We report ROC-AUC, Brier, C@1, F_1, $F_{0.5u}$ and their mean.

Approach	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
method1	0.501	0.744	0.501	0.624	0.544	0.583
method2	0.984	0.918	0.907	0.898	0.954	0.932
Baseline Binoculars	0.972	0.957	0.966	0.964	0.965	0.965
Baseline Fast-DetectGPT (Mistral)	0.876	0.8	0.886	0.883	0.883	0.866
Baseline PPMd	0.795	0.798	0.754	0.753	0.749	0.77
Baseline Unmasking	0.697	0.774	0.691	0.658	0.666	0.697
Baseline Fast-DetectGPT	0.668	0.776	0.695	0.69	0.691	0.704
95-th quantile	0.994	0.987	0.989	0.989	0.989	0.990
75-th quantile	0.969	0.925	0.950	0.933	0.939	0.941
Median	0.909	0.890	0.887	0.871	0.867	0.889
25-th quantile	0.701	0.768	0.683	0.657	0.670	0.689
Min	0.131	0.265	0.005	0.006	0.007	0.224

Table 2
Overview of the mean accuracy over 9 variants of the test set. We report the minumum, median, the maximum, the 25-th, and the 75-th quantile, of the mean per the 9 datasets.

Approach	Minimum	25-th Quantile	Median	75-th Quantile	Max
method1	0.513	0.561	0.571	0.582	0.583
method2	0.527	0.896	0.922	0.926	0.947
Baseline Binoculars	0.342	0.818	0.844	0.965	0.996
Baseline Fast-DetectGPT (Mistral)	0.095	0.793	0.842	0.931	0.958
Baseline PPMd	0.270	0.546	0.750	0.770	0.863
Baseline Unmasking	0.250	0.662	0.696	0.697	0.762
Baseline Fast-DetectGPT	0.159	0.579	0.704	0.719	0.982
95-th quantile	0.863	0.971	0.978	0.990	1.000
75-th quantile	0.758	0.865	0.933	0.959	0.991
Median	0.605	0.645	0.875	0.889	0.936
25-th quantile	0.353	0.496	0.658	0.675	0.711
Min	0.015	0.038	0.231	0.244	0.252

3.3. Conclusion

In this paper, we have completed the tasks set by PAN and have employed the next-token prediction method to tackle the Generative AI Authorship Verification task. Instead of using fine-tuned transformer-based method techniques, we utilize the next-token prediction method to narrow the gap between downstream tasks and pre-training tasks. Finally, on the test set, our performance metrics at the Minimum, 25-th Quantile, Median, 75-th Quantile, and Max were 0.527, 0.896, 0.922, 0.926, and 0.947 respectively. These results certify the effectiveness of our proposed method in performing the Generative AI Authorship Verification task.

Limitations

Firstly, the method proposed in this paper does not involve any prompts in the current LLMs-generated text detection task. Using prompts can better leverage the internal knowledge of language models. Therefore, in future work, we plan to incorporate prompts to complete this task.
Additionally, transforming the task into a binary AI detection task, rather than judging which AI -authored the text, is another method to accomplish AI detection tasks. However, this approach can easily lead to data imbalance issues, where the amount of human-authored data is not equivalent to that of AI-generated data. To address this, data augmentation techniques could be employed to increase the quantity of human-authored data.

Acknowledgments

This research was supported by the National Social Science Foundation of China (22BTQ101)

References

[1] Y. Chen, H. Kang, V. Zhai, L. Li, R. Singh, B. Raj, Token prediction as implicit classification to identify llm-generated text, arXiv preprint arXiv:2311.08723 (2023).
[2] Z. Lai, X. Zhang, S. Chen, Adaptive ensembles of fine-tuned transformers for llm-generated text detection, arXiv preprint arXiv:2403.13335 (2024).
[3] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.
[4] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, Y. Yang, Longt5: Efficient text-to-text transformer for long sequences, arXiv preprint arXiv:2112.07916 (2021).
[5] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.
[6] J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, B. Stein, Overview of the Voight-Kampff Generative AI Authorship Verification Task at PAN 2024, in: G. F. N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[8] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.