Enhancing Human-Machine Authorship Discrimination in Generative AI Verification Task with BERT and Augmented Data

Notebook for PAN at CLEF 2024

Haojie Cao, Zhongyuan Han*, Jingyan Ye, Biao Liu and Yong Han

Foshan University, Foshan, China

Abstract

Voight-Kampff Generative AI Authorship Verification, a task proposed jointly by PAN and the ELOQUENT Lab [1], aims to differentiate between human and machine authors by analyzing text features, addressing the escalating text generation abilities of large language models. This paper uses the BERT pre-trained model for binary classification, fine-tuning it to identify distinctive features of human and machine writing. Additionally, we introduced an augmented dataset to enhance the model's recognition capabilities. Through comprehensive evaluation, we achieved a ranking score of 0.778 on all test sets and achieved 11th place.

Keywords

PAN 2024, Voight-Kampff Generative AI Authorship Verification, BERT, Text classification

1. Introduction

With the rapid advancement and increasing adoption of Large Language Models (LLMs), distinguishing whether a given text is authored by a human or a machine has become increasingly challenging. The focus of "Voight-Kampff Generative AI Authorship Verification 2024" [2] is on how to differentiate between human and machine authors within text.

We propose a binary classifier based on the BERT model to address this task [3]. The model, pre-trained using BERT, assesses the likelihood that a text is authored by a human. To further enhance the predictive accuracy of the model, additional datasets are introduced for data augmentation. These approach aim to further alleviate any data imbalance issues and improve the overall predictive accuracy of the model.

2. Approach

The approach consists of three steps: 1) constructing the dataset, 2) fine-tuning the BERT model, and 3) building a classifier to obtain the results. The PAN dataset provides articles written by humans and generated by AI. Human-written articles are labeled as 1 and AI-generated texts as 0. Additionally, the dataset is augmented by collecting other human-written texts from Hugging Face. This enhanced dataset is then used to fine-tune the BERT model.

To prevent catastrophic forgetting that could render the PAN dataset ineffective within the model, the model is initially trained using the augmented dataset. Subsequently, fine-tuning is performed using the dataset provided by PAN to ensure the model retains and effectively utilizes the specific characteristics of the PAN data. Finally, the output from BERT is fed into a simple classifier to perform binary classification. This classifier produces a result of either 0 or 1, indicating whether the text is generated by a human or an AI.

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

caohaojie0322@163.com (H. Cao); hanzhongyuan@gmail.com (Z. Han*); yyyxy0604@163.com (J. Ye); hyticen@gmail.com (B. Liu); hanyong2005@fosu.edu.cn (Y. Han)

0000-0002-8365-168X (H. Cao); 0000-0001-8960-9872 (Z. Han*); 0009-0002-2749-9422 (J. Ye); 0009-0000-3031-9758 (B. Liu); 0000-0002-9416-2398 (Y. Han)

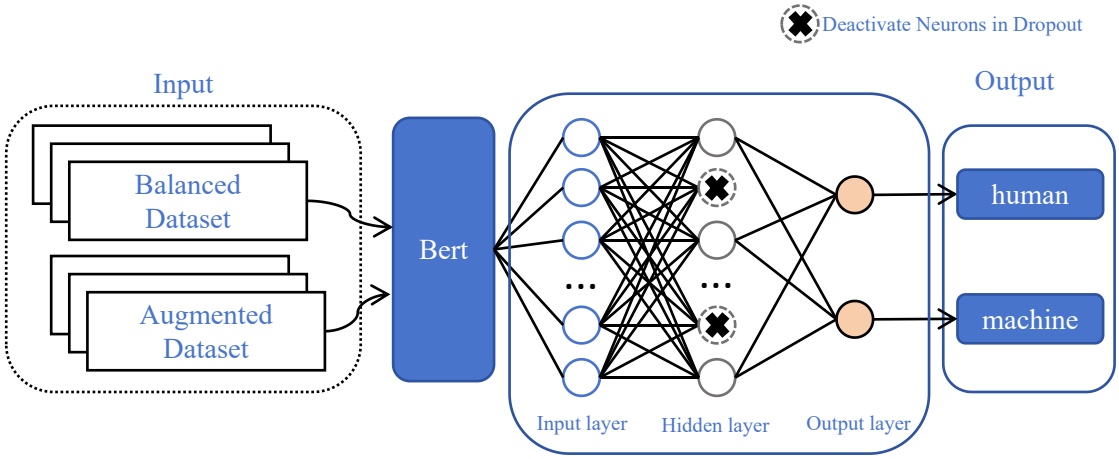

Figure 1: Architecture of the binary classifier based on the BERT model:A fully connected layer is added after BERT, and Dropout is used to prevent overfitting.

2.1. Building Dataset

The dataset provided by PAN comprises over a thousand human-written texts, alongside texts generated by 13 distinct AI models. Each AI model generated an equal number of texts as the human-written ones. To create a balanced dataset, one-thirteenth of the texts were randomly sampled from each of the 13 sets of generative AI texts. This approach allowed us to construct a dataset where the number of human-written texts is equal to the number of generative AI texts, ensuring a 1:1 ratio between human and machine-generated content. The remaining generative AI texts were not discarded. Instead,

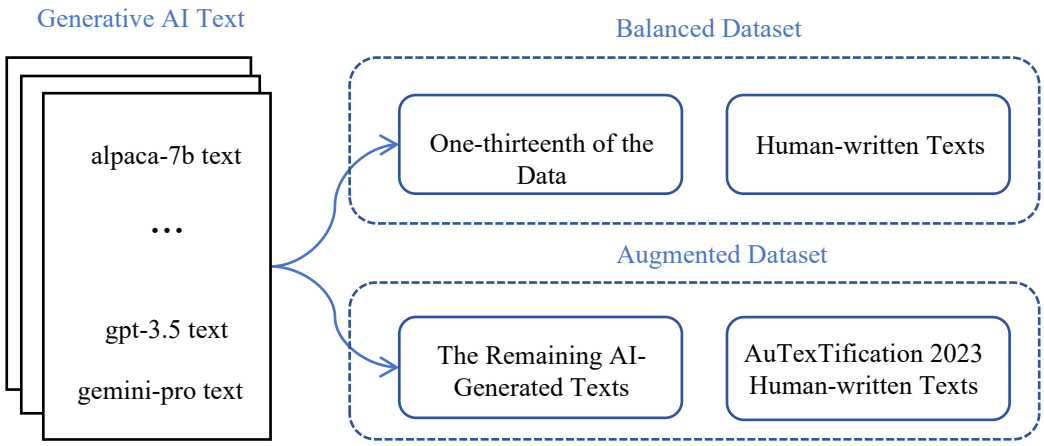

Figure 2: The composition of the dataset

they were supplemented with an equal number of human-written texts collected from Hugging Face's Dataset Card for AuTexTification 2023 [4]. These texts were combined to form an augmented dataset, also referred to as the augmented dataset. This additional dataset ensures a broader representation of language patterns and helps improve the model's ability to generalize. Finally, the Balanced Dataset and Augmented Dataset will be randomly divided into training and validation sets, respectively, with the training set accounting for 70% of the data and the validation set accounting for 30%.

2.2. Fine-tuning the BERT Model

During fine-tuning, no parameters of the base BERT model were fixed. All parameters were updated to adapt to the binary classification task. Upon acquiring the BERT model, a fully connected (Dense) layer

was added to map the model's output to the label space of the binary classification task. This Dense layer has two output units.

To prevent overfitting, Dropout was applied to this fully connected layer, randomly setting a portion of the neurons' outputs to zero during each training iteration. This step helps increase the model's generalization capability. The Dropout rate was set to 0.1, meaning that 10% of the neurons are dropped out. Additionally, a softmax activation function was used to convert the model's outputs into class probabilities. To train the model, the loss function, optimizer, and evaluation metrics were configured. Considering that the labels of the task are integer class identifiers, sparse categorical cross-entropy was selected as the loss function. The optimizer chosen was Adam [5], with a learning rate set to 2e-5. This relatively small learning rate helps stabilize the training process and prevent gradient explosion. The evaluation metric used was accuracy, which measures the model's performance in the classification task.To prevent catastrophic forgetting, the model was first trained on the augmented dataset for ten epochs. After each epoch, its performance was evaluated on the validation set to select the best-performing model. Subsequently, the model was trained using the balanced dataset.

2.3. Build Classifier

In TIRA [6], the prediction task involves evaluating pairs of sentences, where the model needs to determine which sentence is written by a human. To address this task, a simple classifier was developed to compare the probabilities of the two sentences being written by a human. The classifier outputs 0 if the probability of the first sentence being human-written is greater, and outputs 1 if the probability of the second sentence being human-written is greater. This straightforward approach enables the model to make binary predictions based on the relative likelihood of human authorship for each sentence pair.

3. Result

Based on the aforementioned methods, we identified the texts in the competition and uploaded our results. Table 1 presents the final scores of the Voight-Kampff Generative AI Authorship Verification 2024 shared task, where the individual validity score is an aggregate across all test datasets, corrected by half a standard deviation to penalize unstable classification performance. Rankings are based on the mean average of all individual scores. Our team secured the 11th position out of 30, achieving a score of 0.778 across all test datasets. Table 2 provides an overview of the accuracy in detecting whether a text is written by a human in Task 4 of PAN 2024 (Voight-Kampff Generative AI Authorship Verification). Our model achieved a mean score of 0.906, surpassing most of the published baselines. In addition to the primary test dataset, the PAN organizers evaluated the "Voight-Kampff" Generative AI Authorship Verification on nine additional variants. Table 3 showcases the overview of the mean accuracy across these nine variants of the test set. Among the nine variant datasets, our model demonstrated its lowest accuracy at 0.361, while achieving a 75th percentile score of 0.959, with the highest accuracy reaching 1. Our model consistently outperformed the baselines across most of these variant datasets, underscoring its robustness and effectiveness in diverse scenarios.

Table 1
Overview of the comprehensive results and rankings across all test sets on PAN 2024 (Voight-Kampff Generative AI Authorship Verification). Report ROC-AUC, Brier, C@1, F_1, $F_{0.5u}$ and their mean.

Rank	Team	SYSTEM	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
11	heartsteel(Our)	canary-paint	0.799	0.788	0.740	0.740	0.741	0.763
	Baseline Binoculars (Falcon-7B)		0.751	0.780	0.734	0.720	0.720	0.741
	Baseline DetectLLM-LRR (Mistral-7B)		0.656	0.758	0.617	0.618	0.618	0.654
	Baseline Fast-DetectGPT (Mistral-7B)		0.637	0.710	0.616	0.611	0.608	0.638
	Baseline Text Length		0.608	0.607	0.607	0.596	0.596	0.604

Table 2
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
canary-paint(Our)	0.906	0.906	0.906	0.906	0.907	0.906
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 3
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
canary-paint(Our)	0.361	0.791	0.901	0.959	1.000
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

4. Conclusion

In this study, we addressed the challenge of distinguishing between human and machine-generated texts in the Voight-Kampff Generative AI Authorship Verification 2024 task. By leveraging a BERT-based model and incorporating data augmentation techniques, we enhanced the model's ability to accurately classify texts. Our approach involved constructing a balanced dataset, fine-tuning the BERT model, and implementing a simple classifier for binary predictions.

The evaluation results demonstrated the feasibility and effectiveness of our approach, achieving a mean score of 0.778 and an overall ranking of 11, which outperformed all baseline models. These findings suggest that our approach is a promising solution for human-machine authorship discrimination, contributing to the broader field of AI authorship verification.

Acknowledgements

This work is supported by The Natural Science Foundation of Guangdong Province, China (No.2022A1515011544)

References

[1] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, 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.

[2] J. Bevendorff, M. Wiegmann, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, E. Stamatatos, M. Potthast, B. Stein, Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[3] M. Fabien, E. Villatoro-Tello, P. Motlicek, S. Parida, Bertaa: Bert fine-tuning for authorship attribution, in: Proceedings of the 17th International Conference on Natural Language Processing (ICON), 2020, pp. 127–137.

[4] A. M. Sarvazyan, J. Á. González, M. Franco-Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains, in: Procesamiento del Lenguaje Natural, Jaén, Spain, 2023.

[5] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[6] 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.