BertT: A Hybrid Neural Network Model for Generative AI Authorship Verification

Notebook for PAN at CLEF 2024

Zepeng Wu, Wenyin Yang* , Li Ma and Zikai Zhao

Foshan University, Foshan, China

Abstract

With the rapid development and widespread adoption of Large Language Models (LLMs), distinguishing between human-authored and machine-generated texts has become increasingly complex. Although various classification methods have been devised to help identify the origins of texts, they often fail to address the fundamental feasibility and inherent challenges of the task. Building on extensive experience in the field of authorship verification, this study introduces BertT, a novel hybrid model that combines BERT and Transformer technologies, specifically designed for the Generative AI Authorship Verification Task organized in collaboration with PAN and ELOQUENT Labs. This task requires accurately identifying human-authored texts from pairs, one written by a human and the other generated by a machine. Leveraging the deep semantic understanding capabilities of BERT and the efficient sequence processing power of Transformers, our model, BertT, significantly outperforms existing baseline models such as Fast-Detect.

Keywords

PAN 2024, Generative AI Authorship Verification, BERT, Transformer

1. Introduction

Text classification is a cornerstone of Natural Language Processing (NLP), with authorship verification serving as a pivotal application in this domain. This process is crucial for validating the authenticity of documents, detecting plagiarism, and identifying the origins of articles, thereby preserving the integrity of written content across various fields. The Generative AI Authorship Verification Task at PAN@CLEF 2024 [1], which builds upon previous challenges, aims specifically to differentiate between human-authored and machine-generated texts. This task is increasingly pertinent as Large Language Models (LLMs) like GPTs now produce high-quality text that closely mimics human writing, thereby presenting substantial challenges in differentiation.

The utility of authorship verification has been demonstrated in various contexts, underscoring its adaptability and critical importance. For example, Halvani et al. explore the use of compression models for authorship verification, highlighting their effectiveness in digital text forensics without relying on complex machine learning algorithms or extensive feature engineering [2]. This approach is well-aligned with our need for efficient and scalable solutions to manage the vast amounts of text generated by LLMs. Similarly, Bevendorff et al. have adapted the unmasking method to short texts, significantly reducing the amount of material required for effective authorship verification, thereby making it applicable to more practical scenarios [3]. Additionally, the challenge of distinguishing between machine-generated and human-authored content is accentuated in the work by Bao et al., who developed Fast-DetectGPT. This model improves the efficiency of detecting machine-generated text through the innovative use of conditional probability curvature, thereby reducing computational costs while maintaining high accuracy [4].

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

2112203031@stu.fosu.edu.cn (Z. Wu); cswyyang@fosu.edu.cn (W. Yang); molly_917@163.com (L. Ma) ; gzjbzzk@163.com (Z. Zhao)

0009-0004-5756-9713 (Z. Wu); 0000-0003-4842-9060 (W. Yang); 0000-0002-5013-052X (L. Ma) ; 0009-0006-7120-3958 (Z. Zhao)

This advancement is particularly relevant to our study as it addresses similar challenges concerning processing efficiency and accuracy. Moreover, the broad application of neural networks in text classification tasks is exemplified by Yang et al. and Yuan et al. in their respective studies on profiling irony and stereotype spreaders on Twitter. These studies employ RNN and CNN models to classify complex social media content, offering insights into the adaptability of neural network architectures for varied NLP tasks [5], [6].

In response to the complexities of authorship verification in the era of LLMs, we developed Bert_T, a novel model that marries the deep semantic understanding capabilities of BERT with the efficient sequential data processing power of Transformer architecture. This model employs a sophisticated contrastive learning approach with an advanced loss function, aimed at enhancing the discrimination between human and machine text. It operates on a dataset formatted in pairs "(text1, text2, label)", training to detect subtle nuances that signify distinct authorship styles. Prior to its effectiveness evaluation, we submitted our model to the TIRA.io platform [7], which provides a stringent and controlled testing environment for fair and transparent benchmarking against established baselines. This preliminary submission was crucial for assessing the model's real-world applicability and refining its performance based on unbiased feedback. The Bert_T demonstrated superior performance across several key metrics, achieving a ROC-AUC of 0.967, a Brier score of 0.903, a C@1 of 0.869, an F1 score of 0.869, and an F0.5u of 0.872, culminating in an overall mean score of 0.896. These results significantly surpassed those of other baseline models such as Fast-DetectGPT (Mistral), PPMd, Unmasking, and Fast-DetectGPT, underscoring the Bert_T's enhanced ability to discern between human and machine-generated texts. This success highlights the efficacy of our approach in tackling the complexities of generative AI authorship verification.

2. Dataset

The dataset for the Generative AI Authorship Verification Task at PAN@CLEF 2024 plays a crucial role in training and validating the efficacy of our Bert_T model. This year, the dataset comprises a diverse array of text genres, reflecting a mix of both real and synthetically generated content. The primary sources of data include news articles, Wikipedia introduction texts, and pieces of fanfiction, which provide a rich variety in style, structure, and complexity. Additionally, PAN participants receive a bootstrap dataset that includes real and fabricated news articles covering various 2021 U.S. news headlines, designed to simulate scenarios that models might encounter in practical applications.

The data, sourced from contributions by ELOQUENT participants, is meticulously curated to ensure a balanced representation of human and machine-authored texts. The bootstrap dataset is formatted as newline-delimited JSON files, where each file contains a list of articles. These articles are authored either by one or more human authors or entirely by an AI, specifically Google's Gemini Pro model. The dataset structure is pivotal for the task, as it contains pairs of texts where each pair is written on the same topic but by different authors-one human and one machine. The file format for these pairs is demonstrated below:

{"id": "gemini-pro/news-2021-01-01-2021-12-31-kabulairportattack/art-081", "text": "..."}
{"id": "gemini-pro/news-2021-01-01-2021-12-31-capitolriot/art-050", "text": "..."}

Each text pair in the dataset is meticulously labeled with `0` or `1`, indicating whether the texts are from the same author, thereby facilitating supervised learning. The test dataset is provided in a slightly altered format to challenge the model's ability to generalize. Instead of individual files, it is delivered as a single JSONL file where each line contains a pair of texts. The content of this file is arranged such that the identities of the authors are anonymized, and the order of texts scrambled:

{"id": "iixcWBmKWQqLAwVXxXGBGg", "text1": "...", "text2": "..."}
{"id": "y12zUebGVHSN9yiL8oRZ8Q", "text1": "...", "text2": "..."}

Participants are tasked with predicting which of the two texts in each pair is human-authored. This setup tests the model's ability to discern subtle linguistic and stylistic nuances that typically

distinguish human writing from its AI-generated counterpart. Access to the dataset is regulated via Zenodo, where participants must register and request access using their Tira-registered email, ensuring that the use of this data remains confined to research purposes and that no redistribution occurs. This controlled distribution ensures compliance with copyright regulations and maintains the integrity of the data for academic and developmental uses.

3. Methodology

3.1. Dataset Preprocessing

Effective data preprocessing is essential for the robust performance of machine learning models, particularly in tasks involving natural language processing such as authorship verification. For the Generative AI Authorship Verification Task at PAN@CLEF 2024, our preprocessing routine involved several critical steps to enhance the quality and consistency of model inputs.

Initially, the text normalization process involved removing all punctuation and converting text to lowercase to reduce variability and focus the model's learning on substantive content. This was coupled with the removal of non-alphabetic characters and numerals to ensure that the model trained strictly on textual elements. Following normalization, stopwords-common words that typically do not contribute to the identification of authorship-were removed to minimize data noise and enhance focus on more distinctive text features.After cleaning the texts, the corpus was tokenized into individual words or tokens, which is essential for structuring raw text into a format suitable for machine learning models. The texts were then vectorized using a pre-trained Bert tokenizer, which also standardized the length of tokens through padding and truncation to optimize computational efficiency. To address the challenge of limited training data, we implemented data augmentation techniques to artificially expand the dataset, creating new text pairs from existing ones by subtly modifying texts while preserving their key attributes. This approach helped improve the model's generalization capabilities from training scenarios to real-world applications.

Throughout the preprocessing stages, we meticulously ensured that the alterations did not compromise the semantic integrity or the stylistic attributes of the texts, which are crucial for authorship identification. This comprehensive preprocessing not only prepared the dataset for effective training of our Bert_T model but also enhanced the model's accuracy in distinguishing between human and machine-generated texts, a critical aspect of the verification task.

3.2. Network Architecture

In this study, we introduced Bert_T, a hybrid neural network model that integrates BERT-base for robust feature extraction with a Transformer encoder to handle attention-based dynamics, specifically tailored for distinguishing between human-written and machine-generated texts. We employ the bert-base-uncased model from Hugging Face's Transformers library as our foundational pre-trained BERT layer, leveraging its well-established capabilities in natural language understanding. This layer focuses on the CLS token embedding to capture comprehensive textual context, which is then processed through a Dropout layer to prevent overfitting and enhance generalizability. The Transformer Encoder, equipped with a multi-head attention mechanism, dynamically integrates information across text segments, crucial for identifying subtle linguistic and stylistic nuances. During testing, Bert_T processes each text in a pair independently in a JSONL format, evaluating the likelihood of each text being human-written and comparing these scores to classify texts; the text with the higher score is deemed human-authored. Optimization of model parameters such as learning rate and batch size, along with the use of Binary Cross-Entropy Loss, fine-tunes the model's accuracy, ensuring it performs effectively on metrics such as ROC-AUC and Brier scores. This configuration enables Bert_T to meet the specific challenges of the Generative AI Authorship Verification Task at PAN@CLEF

2024, demonstrating both innovative theoretical approaches and practical discriminative capabilities, as illustrated in Figure 1: Bert_T Architecture.

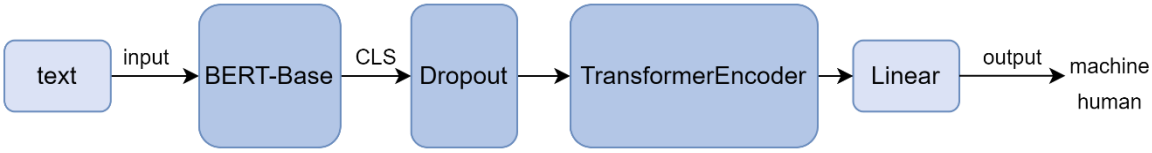

Figure 1: Bert_T Architecture

4. Experiments and Results

4.1. Experimental Setting

In our experimental setup for evaluating the Bert_T model's ability to distinguish between human-authored and machine-generated texts, we preprocessed the dataset and divided it into training and testing sets with a 7:3 ratio. The model, a Bert_T, integrates a pretrained BERT base model with a Transformer layer tailored for sequence classification, featuring 768 hidden units, four attention heads, and a linear classifier. Training parameters were meticulously set, with a batch size of 8 and a learning rate of 1e-6 over 300 epochs using the AdamW optimizer on CUDA-capable GPUs to balance computational efficiency and learning depth.

4.2. Metrics

Our evaluation framework was meticulously designed to rigorously assess the performance of the Bert_T model across several metrics that reflect its effectiveness in distinguishing between human-authored and machine-generated texts. The model was evaluated using a standard set of metrics that are commonly employed in authorship verification tasks, including ROC-AUC, Brier score, C@1, F1, and F0.5u, along with the arithmetic mean of these metrics to provide a comprehensive overview of performance.

Performance Metrics:

ROC-AUC measures the area under the receiver operating characteristic curve, providing insight into the model's ability to discriminate between classes across all thresholds [8]. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The formula is given by:

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(t)\ d\big(\text{FPR}(t)\big) \tag{1}$$

Brier Score evaluates the mean squared error of the probabilities assigned, indicating the accuracy of probability predictions [9]. The lower the Brier score, the better, as it reflects a closer proximity to the true outcome. It is calculated as:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(\text{predicted probability}_i - \text{actual outcome}_i)^2 \tag{2}$$

C@1 represents a modified accuracy that treats non-answers (predictions with a confidence score of 0.5) by averaging the accuracy of the remaining cases, thus penalizing uncertainty [10]. This metric is particularly useful in situations where making no prediction is preferable to making an incorrect prediction. The formula is:

$$C@1 = \frac{\text{Number of correct answers}}{\text{Total number of cases} - \text{Number of non-answers}} + \frac{\text{Number of non-answers}}{\text{Total number of cases}} \tag{3}$$

F1 Score is the harmonic mean of precision and recall, offering a balance between the precision of the classifier and its recall capability [11]. It is particularly useful in situations where an equal balance between precision and recall is desired. The formula is:

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Where Precision $= \frac{\text{TP}}{\text{TP+FP}}$ *and* Recall $= \frac{\text{TP}}{\text{TP+FN}}$.

F0.5u is a variant of the F-measure that weights precision more than recall, suitable for scenarios where false positives are more costly than false negatives [12]. It is calculated using the formula:

$$F0.5u = (1 + 0.5^2) \cdot \frac{\text{Precision} \times \text{Recall}}{0.5^2 \cdot \text{Precision} + \text{Recall}} \quad (5)$$

These metrics collectively provided a robust framework for evaluating our model, enabling us to effectively measure its ability to perform authorship verification across different dimensions of accuracy and reliability.

4.3. Results

Our Bert_T model demonstrated robust performance in the PAN 2024 Voight-Kampff Generative AI Authorship Verification task, showcasing substantial effectiveness across several critical metrics. As evidenced in Table 1, Bert_T achieved a ROC-AUC of 0.967, which, while slightly lower than the top-performing Baseline Binoculars at 0.972, reflects a high level of discriminative capability. The Brier score for Bert_T was 0.903, indicating reliable probability predictions of class membership, although it did not surpass the Baseline Binoculars, which scored 0.957. Regarding precision-related metrics, Bert_T recorded scores of 0.869 for both C@1 and F1, and 0.872 for F0.5u, remaining competitive although below the near-perfect scores around the 95th percentile.

Table 2 presents an overview of Bert_T's mean accuracy across nine test set variants, showing considerable stability and less variability in performance compared to other models which displayed more significant fluctuations. Bert_T maintained a minimum accuracy of 0.354 and a maximum of 0.980, with a notable median performance of 0.892, and the 25th and 75th percentiles at 0.864 and 0.896, respectively. These figures underscore Bert_T's robust performance across different testing scenarios, highlighting its efficacy in handling the complex demands of the verification task.

In terms of competition standings, our submission ranked 20th out of 30 participants on the official PAN 2024 leaderboard. Notably, Bert_T outperformed all but one baseline with a ranking score over all test datasets of 0.608, as detailed on the PAN 2024 leaderboard. This ranking underscores our model's competitive edge and its significant discriminative power in a challenging environment filled with diverse and sophisticated entries.

These results affirm that Bert_T not only embodies theoretical innovation but also exhibits significant practical capabilities in the authorship verification domain. The model's ability to effectively discern between human and machine-generated texts makes it a valuable tool for complex text analysis tasks. Future work will focus on further optimizing model parameters, enhancing feature engineering techniques, and expanding the diversity of the training dataset to boost the model's generalizability and performance across varied textual contexts. This continuous improvement aims to refine Bert_T's capabilities for higher detection accuracy and broader application scope in real-world scenarios.

Table 1:The final performance of our submission on PAN 2024 (Voight-Kampff Generative AI Authorship Verification)

Approach	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
Bert_T	0.967	0.903	0.869	0.869	0.872	0.896
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

Approach	Minimum	25-th Quantile	Median	75-th Quantile	Max
Bert_T	0.354	0.864	0.892	0.896	0.980
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

5. Conclusion

This paper details the development and evaluation of the Bert_T model, our innovative contribution to the PAN 2024 Voight-Kampff Generative AI Authorship Verification task. Combining BERT-based feature extraction with a Transformer encoder for attention processing, Bert_T effectively differentiates between human-written and machine-generated texts. It demonstrated strong performance across various metrics, achieving a ROC-AUC of 0.967 and a Brier score of 0.903, which confirms its reliability in predictions. Despite stiff competition from established baselines, Bert_T maintained consistent performance across different test set variants, with accuracies ranging from a minimum of 0.354 to a maximum of 0.980. This showcases its capability to handle diverse and complex textual scenarios effectively. Moving forward, we plan to further refine Bert_T by optimizing its parameters, enhancing its feature engineering techniques, and expanding its training dataset to cover a broader spectrum of text types and genres. These efforts will not only improve the model's performance in authorship verification tasks but also extend its applicability to a wider range of natural language processing challenges, aiming for higher detection accuracy and broader operational scope.

6. Acknowledgements

This work was supported by grants from the Guangdong-Foshan Joint Fund Project (No. 2022A1515140096) and Open Fund for Key Laboratory of Food Intelligent Manufacturing in Guangdong Province (No. GPKLIFM-KF-202305).

References

[1] 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International

Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.
[2] O. Halvani, C. Winter, and L. Graner. "On the usefulness of compression models for authorship verification." Proceedings of the 12th international conference on availability, reliability and security. 2017.
[3] J. Bevendorff, B. Stein, M. Hagen, et al. "Generalizing unmasking for short texts." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019.
[4] G. S. Bao, Y. B. Zhao, Z. Y. Teng, et al. "Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature." arXiv preprint arXiv:2310.05130 (2023).
[5] Z. X. Yang, L. Ma, W. Y. Yang , et al. A Intelligent Detection Method for Irony and Stereotype Based on Hybird Neural Networks. In Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast, editors, CLEF 2022 Labs and Workshops, Notebook Papers, September 2022. CEUR-WS.org.
[6] D. Yuan, W. Y. Yang, L. Ma, et al. Analysis of Irony and Stereotype Spreaders Based On Convolutional Neural Networks. In Guglielmo Faggioli, Nicola Ferro, Allan Hanbury, and Martin Potthast, editors, CLEF 2022 Labs and Workshops, Notebook Papers, September 2022. CEUR-WS.org.
[7] 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. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.
[8] A. M. Carrington, D. G. Manuel, P. W. Fieguth, et al. "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.1 (2022): 329-341.
[9] W. Yang, J. Jiang, E. M. Schnellinger, et al. "Modified Brier score for evaluating prediction accuracy for binary outcomes." Statistical methods in medical research 31.12 (2022): 2287-2296.
[10] A. Peñas, A. Rodrigo, A simple measure to assess non-response (2011).
[11] F. Pedregosa, G. Varoquaux, A. Gramfort, Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.
[12] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing unmasking for short texts, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 654–659.