AI Authorship Verification Based On Deberta Model

Notebook for the PAN Lab at CLEF 2024

Ye Zhu, Leilei Kong†

Foshan University, Foshan, Guangdong, China

Abstract

Generative AI Authorship Verification is the task of distinguishing between human-authored and machine-generated texts. This paper explores the application of the pre-trained language model Deberta to address this problem. Our approach involves fine-tuning the Deberta model on a curated dataset comprising both human and machine-generated texts. To manage the imbalance in our dataset, we employed a random sampling to ensure a balanced representation of both types of texts during training. Preliminary experiments show that while our method performs comparably with existing approaches, there is significant potential for further optimization and improvement in identifying human-authored texts. Future work will explore advanced techniques and larger datasets to enhance model.

Keywords

Authorship Verification, Machine-generated Texts, Deberta

1. Introduction

With Large Language Models (LLMs) improving at breakneck speed and seeing more widespread adoption every day, it is getting increasingly hard to discern whether a given text was authored by a human being or a machine. These models, such as GPT-3[1], GPT-4[2], and others, generate text that is often indistinguishable from human writing, posing significant challenges for various applications, including academic integrity, content verification, and online misinformation.

Many classification approaches have been devised to help humans distinguish between human and machine-authored text. Traditional methods rely on surface-level features such as word frequency, syntactic patterns, and stylistic elements, but these features can be easily mimicked by advanced LLMs[3, 4]. Thus, the task of authorship verification in the context of human vs. machine text remains a critical and challenging problem.

Recently, PAN 2024[5] posed a task: given two texts, one written by a human and the other by a machine, identify the human-authored text[6]. We approached this as a binary classification task, which simplifies the challenge and allows us to focus on identifying the most distinctive features of human-authored texts.

To address this task, we use the Deberta[7] model as the pre-trained model, an improved version derived from Bert[8], known for its effective text feature encoding. Deberta's architecture enhances the attention mechanism, making it more adept at capturing intricate patterns in the text. However, sometimes the data we get is not as perfect as we expected.Therefore, in situations where datasets are limited and imbalanced, we adopted a random sampling method that efficiently and economically trains the model by selectively sampling portions of the text data. This approach involves randomly selecting a subset of machine-generated samples for each training epoch while including all human-generated samples. By retaining the most relevant features and reducing the computational load, our method optimizes the training process. Previous studies have shown that specific sampling techniques can significantly improve model performance in imbalanced data scenarios[9, 10, 11].

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

† Corresponding author

kwojmjmqa1744@gmail.com (Y. Zhu); kongleilei@fosu.edu.cn (L. Kong)

0009-0001-2658-9445 (Y. Zhu); 0000-0002-4636-3507 (L. Kong)

2. Data Analysis

PAN 2024 provides a guided dataset covering both real and fake news articles from multiple 2021 US news headlines. Each file contains a list of articles, written either by (any number of) human authors or a single machine. Machine text is generated by some large language models such as Gemini Pro[12].

The dataset comprises human-authored text and machine-generated text, with a significant imbalance in the ratio between the two categories (1:13). To address this challenge, we adopted a random sampling approach during model training. Due to the limited data availability and the need to balance computational resources, we trained the classification model for two epochs.

In the first epoch, we randomly selected 1200 samples from the combined machine-generated text to ensure representation from different sources and topics. For the second epoch, we increased the sample size to 3000 to further enrich the training data. All human-authored samples were included in both epochs to maintain a balanced representation of human and machine texts.

Subsequently, in each epoch, we combined these two sets of data into a format where a label corresponds a type of text, classifying the two types of texts separately. We then split the data into training, validation, and test sets in a ratio of 0.95, 0.05, and 0.05, respectively. This partitioning strategy ensured that the model was trained on a diverse range of samples while maintaining sufficient data for evaluation and testing. Finally, we utilized the Deberta-large model architecture for AI authorship verification on the combined dataset.

3. Experiments and Results

3.1. Experiment setup

We utilized the Deberta-large model, which is characterized by a vocabulary size of 50,000, a hidden size of 1024, 24 layers, and a total of 3.03 billion parameters. This model was selected for its disentangled attention mechanism and enhanced masked decoder. The classification model was built using PyTorch, with training conducted using a batch size of 2. We did not set a maximum encoder length, fully leveraging the model's capacity to handle long texts. The AdamW[13] optimizer, with a learning rate of 1e-6, was employed to update the model weights, while cross-entropy[14] was used as the loss function to measure prediction error. The network was trained over 2 epochs to ensure thorough learning without overfitting.

In this study, we just utilized the [CLS] token, which is a standard practice in BERT and its derivative models. The [CLS] token is positioned at the onset of the sequence, serving to aggregate information from the entire input sequence, which is crucial for classification tasks. We employed the [CLS] token based on the default settings as per the model's pre-training, without any modifications to its functionality.

All experiments were conducted on an NVIDIA A800 GPU with 80GB of memory, providing the necessary computational power to handle the large model and extensive dataset. Additionally, data augmentation techniques such as random sampling were applied to enhance the training data diversity, thereby improving the model's generalization ability. Performance metrics, including accuracy, precision, recall, and F1-score, were used to evaluate the model's effectiveness on the test dataset.

3.2. Results

To process each sample, we compare the confidence scores of Text 1 and Text 2. If Text 1's confidence score is higher, the final confidence score is 1 minus Text 1's confidence score; if Text 2's confidence score is higher, the final confidence score is Text 2's confidence score.

Table 1 summarizes the performance of the validation and test sets in this experiment, highlighting high accuracy, precision, recall, and F1 scores.Table 2 shows the summarized results averaged (arithmetic mean) over 10 variants of the test dataset. Each variant uses a different technique to test the robustness of authorship verification approaches, such as switching text encoding, translating text, changing

Table 1
This table summarizes the performance of the Deberta model on the validation and test sets, highlighting its robust accuracy, precision, recall, and F1 score across both datasets.

Metric	Validation Set	Test Set
Accuracy	0.9829	0.9853
Precision	0.9853	0.9898
Recall	0.9692	0.9752
F1	0.9769	0.9821

Table 2
Overview of the accuracy in detecting if a text is written by a human on PAN 2024, Voight-Kampff Generative AI Authorship Verification. We report ROC-AUC, Brier, C@1, F1, F0.5u, and their mean.

Approach	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
beige-limit	0.627	0.660	0.590	0.442	0.433	0.555
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
beige-limit	0.307	0.759	0.845	0.864	0.896
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

the domain, and manual obfuscation. Table 3 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). The evaluations for Table 2 and Table 3 were conducted on the PAN 2024 Generative AI Authorship Verification task training dataset using the TIRA[15] platform. Our method, referred to as "beige-limit" in the tables, is compared against various baselines.

4. Conclusion

This paper addresses Generative AI authorship verification using the Deberta model. The goal was to distinguish between human and machine-authored texts. By employing a pre-trained Deberta-large model and random sampling to manage data imbalance, we conducted a series of experiments to evaluate the model's performance.We ranked 23rd in this task using this method.

Our study indicates that advanced pre-trained language models like Deberta have potential for authorship verification tasks. Future research could explore more efficient training strategies and extend this approach to other domains and languages.

Acknowledgments

This research was supported by the Natural Science Platforms and Projects of Guangdong Province Ordinary Universities (Key Field Special Projects) (No. 2023ZDZX1023)

References

[1] L. Floridi, M. Chiriatti, Gpt-3: Its nature, scope, limits, and consequences, Minds and Machines 30 (2020) 681–694.

[2] J. Achiam, S. Adler, S. Agarwal, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[3] D. Ippolito, D. Duckworth, C. Callison-Burch, et al., Automatic detection of generated text is easiest when humans are fooled, arXiv preprint arXiv:1911.00650 (2019).

[4] R. Zellers, A. Holtzman, H. Rashkin, et al., Defending against neural fake news, Advances in neural information processing systems 32 (2019).

[5] 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.

[6] 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.

[7] P. He, X. Liu, J. Gao, et al., Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[8] J. Devlin, M. W. Chang, K. Lee, et al., Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, et al., Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[10] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent data analysis 6 (2002) 429–449.

[11] M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural networks 106 (2018) 249–259.

[12] T. G, R. Anil, S. Borgeaud, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).
[13] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
[14] R. Rubinstein, The cross-entropy method for combinatorial and continuous optimization, Methodology and computing in applied probability 1 (1999) 127–190.
[15] 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.