Team karami-kheiri at PAN: Enhancing Machine-Generated Text Detection with Ensemble Learning Based on Transformer Models

Notebook for the PAN Lab at CLEF 2024

Mohammad Karami Sheykhlan[1,*], Saleh Kheiri Abdoljabbar[2] and Mona Nouri Mahmoudabad[1]

[1]*University of Mohaghegh Ardabili, Daneshgah St., Ardabil, 5619911367, Iran*

[2]*University of Tabriz, Bahman Boulevard, Tabriz, 5166616471, Iran*

Abstract

The advancement of large language models (LLMs) and their increasing utilization for text generation have raised significant concerns about misinformation and potential misuse. This research focuses on developing a dependable model to differentiate between human-authored text and LLM-generated content. By employing transformer models and ensemble learning techniques, we aim to improve the accuracy and reliability of text classification. Our results indicate that the proposed model effectively identifies and categorizes texts, providing a valuable tool for addressing the challenges posed by the widespread use of LLM-generated text.

Keywords

PAN 2024, Authorship verification, Ensemble learning, Transformers,

1. Introduction

Recent advancements in LLMs, particularly with the development of sophisticated systems like GPT-3.5 and GPT-4, have revolutionized content creation across various fields, from advertising and news writing to education and medical research. These models are now capable of generating text that closely mimics human writing, enhancing productivity in numerous professional workflows. However, this rapid integration comes with significant challenges, including the spread of misinformation [1], ethical dilemmas [2], and academic integrity issues [3, 4, 5]. The ability of LLMs to produce highly convincing but potentially misleading or inaccurate content raises concerns about their misuse in generating fake news, deceptive social media posts, and even facilitating academic dishonesty [6]. As such, it has become increasingly important to develop reliable methods for distinguishing between human-authored and machine-generated texts to mitigate these risks and ensure the responsible use of LLMs. In response to these challenges, PAN@CLEF 2024 has introduced the Voight-Kampff Generative AI Authorship Verification task.

The detection of AI-generated text has become a critical area of research, driven by the need to safeguard the integrity of information across digital platforms. Traditional approaches to text verification, which rely heavily on stylistic and linguistic features, are often insufficient when faced with the sophistication of modern LLMs. These models can generate content that not only mirrors human writing but also adapts to various contexts and styles, making manual and even some automated detection methods obsolete. Consequently, more advanced techniques are required to differentiate between human-authored and machine-generated text effectively.

In this study, initially, we fine-tuned transformer models using the training dataset. This process involved adjusting the model parameters to fit the specific characteristics of our dataset better, thereby improving the model's performance on our particular task. We then compared their accuracy with that of a cumulative learning model. The results indicated that incorporating the cumulative learning model

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

mohammadkaramisheykhlan@gmail.com (M. K. Sheykhlan); salehkheiri@gmail.com (S. K. Abdoljabbar); monanouri.m@gmail.com (M. N. Mahmoudabad)

0000-0003-2316-545X (M. K. Sheykhlan); 0009-0009-3328-7486 (S. K. Abdoljabbar)

improved predictive accuracy. This enhancement highlights the potential of cumulative learning in refining the performance of transformer models and underscores its value in developing more reliable AI systems for text classification tasks.

2. Background

In recent years, the field of AI-text generator detection has seen significant advancements, driven by the increasing sophistication of LLMs such as GPT-3.5 and GPT-4. Early approaches focused on identifying stylistic and lexical discrepancies between human-written and machine-generated texts. These methods often utilized shallow machine learning techniques and handcrafted features, but their effectiveness diminished as LLMs became more advanced. Islam et al. [7] used nine traditional machine learning models and two deep learning models to detect AI-generated texts. They found that the Extremely Randomized Trees model provided better accuracy compared to the other models. In contrast, Prova [8] utilized both traditional models and transformer models for this task and found that the performance of the transformer models was significantly better. Therefore, more recent research has shifted towards leveraging deep learning models, particularly transformers, to improve detection accuracy. Alshammari et al. [9] employed a novel classifier using Transformer-based models AraELECTRA and XLM-R, which significantly outperforms GPTZero and OpenAI Text Classifier, achieving up to 99% accuracy with the integration of a Dediacritization Layer. Nitu and Dascalu [10] introduced a corpus of 60,000 Romanian documents, encompassing both human-written and machine-generated texts across five domains. They present two techniques for detecting machine-generated content: a Transformer-based model and a machine-learning model leveraging linguistic features. The Transformer-based method achieved an F1 score of 0.96, outperforming the linguistic feature-based method in two domains. Additionally, the study includes a text similarity analysis and SHAP analysis to identify key linguistic features influencing the classifier's decisions. Techniques such as fine-tuning pre-trained models on labeled datasets of human and AI-generated text have shown promise. Additionally, ensemble learning methods, which combine multiple model predictions to enhance robustness, have been explored as a way to address the limitations of single-model approaches. Qu and Meng [4] achieved first place in Task 8 of SemEval 2024 by using an ensemble learning approach to detect machine-generated texts. Their innovative method demonstrated superior performance in accurately distinguishing between human-written and AI-generated content, setting a new benchmark in the field.

3. Task and Dataset

With the rapid advancement and widespread adoption of LLMs, distinguishing between human- and machine-authored texts has become increasingly challenging. Leveraging expertise in authorship verification, the Generative AI Authorship Verification Task @ PAN, in collaboration with the Voight-Kampff Task @ ELOQUENT Lab, aims to address this by having participants identify the human-written text from a pair of texts. This builder-breaker setup will see PAN participants develop detection systems, while ELOQUENT participants focus on creating text generation and obfuscation methods.

The dataset includes various genres such as news articles, Wikipedia intro texts, and fanfiction, sourced from ELOQUENT participants. Additionally, a bootstrap dataset containing real and fake news articles from 2021 U.S. headlines is provided for training purposes. To access the dataset, participants must register on TIRA [11] and request access on Zenodo[1]. The dataset contains copyrighted material and is restricted to research use only. The test data will be provided in a single JSONL file format, with each line containing a pair of texts, and participants must identify which text is human-authored.

[1] https://zenodo.org/records/10718757

4. System Overview

4.1. Data preparation

In this task, we focused on improving the classification of text by specifically ignoring newline characters within the data while preserving all other information. This approach allows us to maintain the integrity of the content while addressing the task as a classification problem. To achieve this, we utilized the capabilities of three powerful transformer models: BERT, Roberta, and Electra. Each of these models comes with its own tokenizer, which plays a crucial role in preparing the text for the binary classification process. These models are well-regarded in the field of natural language processing for their effectiveness in understanding and generating human-like text. Considering the inherent 512-token limit of these models, we decided to focus on the first 512 tokens of each text sample. This decision was made to ensure that our models could process the data efficiently without exceeding their maximum token capacity. For text samples that exceeded 512 tokens, only the initial 512 tokens were included in the analysis, while the remaining tokens were disregarded. This approach ensures that we work within the constraints of the models while still capturing a significant portion of the text for classification purposes.

4.2. Transformer-based Models

BERT (Bidirectional Encoder Representations from Transformers), Roberta (Robustly Optimized BERT Approach), and Electra (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) are three advanced transformer models that have significantly impacted the field of NLP. BERT, introduced by Google, revolutionized NLP by utilizing bidirectional training, allowing the model to understand the context of a word based on all its surroundings, leading to state-of-the-art performance on various tasks. Roberta, a refinement of BERT by Facebook AI, enhanced the pre-training process by using larger batches, more data, and longer training times, which resulted in improved performance and robustness. Electra, also developed by Google, introduced a novel pre-training approach where the model learns to distinguish real input tokens from "fake" ones generated by a separate generator model. This method makes Electra more efficient and often faster to train compared to BERT, while still achieving competitive or superior performance. Each of these models has its own tokenizer to prepare text for analysis. These models apply to a variety of NLP tasks, ranging from text classification tasks [12] to text summarization [13] and other tasks.

4.3. Ensemble learning

Hard Ensemble Learning is an advanced technique in machine learning that involves combining multiple models, referred to as base learners, to create a more powerful and robust predictive model. Unlike conventional ensemble methods like bagging and boosting, which mainly aim to combine diverse but weaker models, Hard Ensemble Learning incorporates several complex and high-performance models to address difficult tasks. By harnessing the collective intelligence of various models, each trained on distinct data aspects, this approach enhances overall predictive accuracy and generalization.

In this study, we developed three transformer models: BERT, RoBERTa, and Electra. Each model was fine-tuned using the respective training dataset. Our objective with ensemble learning is to combine the strengths of these models to improve overall system performance. Detailed specifics will be provided in the following section.

5. Experiments

5.1. Hyperparameter tuning and Evaluation

In this study, we utilized Google Colaboratory's GPU to fine-tune the BERT, Roberta, and Electra models. Due to token constraints, we limited our consideration to 512 tokens. For all models, we set the

Table 1
Evaluation measures on the test set. The best result is given in bold.

Model	ROC-AUC	Brier	$C@1$	F_1	$F_0.5u$	Mean
BERT	96.5	99.4	97.8	96	99.4	97.8
Roberta	95.4	99.3	97.6	94.3	99.3	97.3
Electra	94.5	99.2	97.2	93.9	99.2	96.8
Ensemble	**96.8**	**99.5**	**98.3**	**96.4**	**99.5**	**98.1**

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**
Our approach	0.959	0.9	0.935	0.858	0.883	0.907
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

hyperparameter learning rate to 2e-5 and the epoch value to 10.

In our study, we evaluated the system using the same metrics as previous PAN authorship verification tasks [14, 15, 16, 17]. These metrics are:

- ROC-AUC: Measures the area under the Receiver Operating Characteristic curve to assess how well the system distinguishes between different classes.
- Brier Score: Evaluates accuracy by looking at the complement of the mean squared loss.
- C@1: A modified accuracy score that assigns a score of 0.5 to non-answers and calculates the average accuracy of the remaining answers.
- F1 Score: The harmonic mean of precision and recall, providing a balance between the two.
- F0.5u: A precision-weighted F measure that treats non-answers as false negatives, focusing more on precision.
- The arithmetic mean of all the metrics above.

5.2. Results

Given LLMs' complexity and human-like text generation, traditional feature extraction methods used in tasks like Author Profiling [18] and Authorship Attribution [19], which previously produced acceptable results, are no longer sufficient. In this section, we present our experimental results on the Voight-Kampff Generative AI Authorship Verification 2024 task using BERT, Roberta, Electra models, and ensemble learning, which leverage the majority vote of the three models for test data samples. The test samples included an ID, text1, and text2. Participants were required to identify which of text1 or text2 was authored by a human or a machine using numerical values in the is_human variable.

We divided the dataset into 60%, 20%, and 20% for the training, validation, and testing phases, respectively. In our approach to classifying test samples, we employed a straightforward method. If the model determined that a human authored text1, we manually assigned a value less than 0.5 to the is_human variable. Conversely, if the model determined that a human authored text2, we assigned a value greater than 0.5 to the is_human variable. In cases where the model was unable to make a decision, we assigned a value of 0.5 to the is_human variable.

Table 1 shows the performance of the approaches used. The experimental results indicate that the hard voting classifier approach is more effective in distinguishing between human-generated and

LLM-generated texts. Therefore, we employed the Ensemble learning model for evaluation on the final test set. The performance of this approach on the final dataset is presented in Table 2.

6. Conclusion

In this study, we addressed the challenge of distinguishing between human-written texts and those generated by LLMs. Our approach leveraged Ensemble learning, incorporating fine-tuned versions of BERT, Roberta, and Electra models. By utilizing these advanced transformer models and applying meticulous fine-tuning on the provided dataset, we were able to enhance the accuracy and reliability of our text classification system. Our findings underscore the potential of Ensemble learning combined with transformer models in advancing the field of authorship verification and enhancing the detection of AI-generated texts.

References

[1] I. Vykopal, M. Pikuliak, I. Srba, R. Moro, D. Macko, M. Bielikova, Disinformation capabilities of large language models, arXiv preprint arXiv:2311.08838 (2023).

[2] J. P. Wahle, T. Ruas, F. Kirstein, B. Gipp, How large language models are transforming machine-paraphrased plagiarism, arXiv preprint arXiv:2210.03568 (2022).

[3] C. Chen, K. Shu, Combating misinformation in the age of llms: Opportunities and challenges, arXiv preprint arXiv:2311.05656 (2023).

[4] X. Qu, X. Meng, Tm-trek at semeval-2024 task 8: Towards llm-based automatic boundary detection for human-machine mixed text, arXiv preprint arXiv:2404.00899 (2024).

[5] C. Stokel-Walker, Ai bot chatgpt writes smart essays-should academics worry?, Nature (2022).

[6] M. Spiegel, D. Macko, Kinit at semeval-2024 task 8: Fine-tuned llms for multilingual machine-generated text detection, arXiv preprint arXiv:2402.13671 (2024).

[7] N. Islam, D. Sutradhar, H. Noor, J. Raya, M. Maisha, D. Farid, Distinguishing human generated text from chatgpt generated text using machine learning. arxiv, arXiv preprint arXiv:2306.01761 (2023).

[8] N. Prova, Detecting ai generated text based on nlp and machine learning approaches, arXiv preprint arXiv:2404.10032 (2024).

[9] H. Alshammari, A. El-Sayed, K. Elleithy, Ai-generated text detector for arabic language using encoder-based transformer architecture, Big Data and Cognitive Computing 8 (2024) 32.

[10] M. Nitu, M. Dascalu, Beyond lexical boundaries: Llm-generated text detection for romanian digital libraries, Future Internet 16 (2024) 41.

[11] 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.

[12] M. K. Sheykhlan, J. Shafi, S. Kosari, Pars-hao: Hate speech and offensive language detection on persian social media using ensemble learning, Authorea Preprints (2023).

[13] Y. Liu, M. Lapata, Text summarization with pretrained encoders, arXiv preprint arXiv:1908.08345 (2019).

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] H. B. Giglou, M. Rahgouy, T. Rahgooy, M. K. Sheykhlan, E. Mohammadzadeh, Author profiling: Bot and gender prediction using a multi-aspect ensemble approach, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_231.pdf.

[19] M. Rahgouy, H. Giglou, T. Rahgooy, M. Sheykhlan, E. Mohammadzadeh, Cross-domain Authorship Attribution: Author Identification using a Multi-Aspect Ensemble Approach, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.